Analyse von Überlebenszeiten bei Tumorpatienten anhand Methoden des Data Minings

Bachelorthesis

von

Fabian Sailer (177375)
geboren am 23. August 1992 in Heilbronn

Abgabedatum: 25.02.2014

Referent: Prof. Dr. Daniel Pfeifer, Hochschule Heilbronn
Korreferent: Dr. Michael Hanselmann, Robert Bosch GmbH
Betreuerin: Monika Pobiruchin, Hochschule Heilbronn

Hochschule Heilbronn
Fakultät für Informatik
Medizinische Informatik Bachelor

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides Statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen (einschließlich elektronischer Quellen) direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts habe ich Unterstützungsleistung von folgenden Personen erhalten:

- Prof. Dr. Daniel Pfeifer
- Prof. Dr. med. Uwe Martens
- Dr. Michael Hanselmann
- Dr. Med. Sylvia Bochum
- Dipl.–Inform. Med. Monika Pobiruchin

Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt und ist auch noch nicht veröffentlicht.

Datum: 25.02.2014

(Fabian Sailer)

Danksagung

Ich danke allen, die mich in den letzten vier Monaten beim Schreiben und Ausarbeiten dieser Thesis unterstützt und begleitet haben. Vor allem möchte ich mich bei Hr. Pfeifer und Fr. Pobiruchin für die intensive Betreuung während der Arbeit und allen Korrekturlesern, die ihre Zeit geopfert haben um meine Fehler zu finden bedanken.

Ich bedanke mich besonders bei Prof. Dr. Uwe Martens und dem gesamten Team der Onkologie des Gesundbrunnens Heilbronn für die Bereitstellung der Daten und die Kooperation während der Arbeit.

Zusammenfassung

Behandlungen von Tumoren zielen in erster Linie auf eine Verlängerung der Überlebenszeit des Patienten ab. Es ist für Ärzte eine Hilfe, wenn zu Beginn der Behandlung die voraussichtliche Überlebenszeit abgeschätzt werden kann. Dies geschieht aktuell oftmals mit Hilfe einer manuellen Einteilung in Risikoklassen. Für diese sind aus Erfahrungswerten typische Überlebenszeiten bekannt.
In Zeiten der zunehmenden Digitalisierung ist es nur logisch den Versuch zu starten, die Klassifizierung automatisch vorzunehmen. In dieser explorativen Grundlagenarbeit werden zwei Data Mining–Verfahren — namentlich „naiver Bayes Klassifikator“ und „k–means Clustering“ — auf ihre Fähigkeit bezüglich der Überlebenszeitprognose hin untersucht. Dazu werden verschiedene Feature Selection Verfahren (Information Gain, Expertenselektion, Forward Selection, Backward Elimination und No Selection) getestet.
Nach dem k–means Clustering können Kaplan–Meier–Kurven der einzelnen Cluster gezeichnet werden. Aus diesen kann eine Prognose der Überlebenszeit abgelesen werden. Der naive Bayes Klassifikator errechnet nach einer (äquifrequenten oder äquidistanten) Diskretisierung der Überlebenszeit für jeden Patienten individuell eine Überlebens–Wahrscheinlichkeits–Verteilung.
Das Training der Data Mining–Verfahren erfolgte auf der Basis von Datensätzen kolorektaler Tumorpatienten des Tumorregisters des Tumorzentrums Heilbronn–Franken.

Inhaltsverzeichnis

1. Einleitung

1.1. Hintergrund

Die SLK–Kliniken Heilbronn GmbH sind der größte Gesundheitsdienstleister der Region Heilbronn–Franken und bilden einen Zusammenschluss mehrerer Krankenhäuser. Das größte der Häuser in diesem Zusammenschluss ist das Klinikum am Gesundbrunnen in Heilbronn. Viele Kliniken innerhalb des Krankenhauses am Gesundbrunnen sind zertifiziert und garantieren daher eine hohe Qualität der Patientenversorgung und betreibt unter anderem ein 2013 re–zertifiziertes Tumorzentrum. Eines der Spezialgebiete dieses Tumorzentrums ist die Behandlung von Kolonkarzinomen. Seit Mitte der 80er Jahre werden alle behandelten Fälle dabei unter anderem in einem eigenen Tumorregister der Onkologie dokumentiert. Bisher wurden allerdings noch keine systematischen Auswertungen auf diesen Daten vorgenommen, was diese Datensätze für informationsverarbeitende Prozesse besonders interessant macht (nach [slk14]).

Diese explorative Arbeit versucht, mittels verschiedener Methoden des Data Minings eine Aussage bezüglich der voraussichtlichen Überlebenszeit von Tumorpatienten zu treffen. Dazu sollen auf vorhandenen — bevorzugt lokalen — Tumorregistern unterschiedliche Verfahren untersucht werden. Durch diese Analysen sollen Faktoren ausfindig gemacht werden, die Aussagen über die Überlebenszeit von Tumorpatienten ermöglichen.

1.2. Bedeutung

Durch medizinischen und wirtschaftlichen Fortschritt erhöhte sich die durchschnittliche Lebenserwartung in Industrienationen in den letzten Jahren konstant. Da Tumorerkrankungen durch degenerative Veränderungen des Genmaterials entstehen, sind vor allem, aber nicht ausschließlich, ältere Menschen von diesen betroffen. Somit stieg in den letzten Jahren auch die Anzahl an Tumorerkrankungen in der Bevölkerung. Inzwischen sind Tumorerkrankungen die dritthäufigste Todesursache in Industrienationen (siehe [RM08]).

Eine Kenntnis der Überlebenszeit, oder zumindest einer Abschätzung davon, der Tumorpatienten, kann für die Behandlung und deren Verlauf von entscheidender Bedeutung sein. Zum einen steht zu Beginn der Behandlung die Frage, ob ein Patient kurativ oder palliativ behandelt werden kann, zum anderen kann nach Abschluss der Behandlung deren Erfolg kontrolliert werden, indem die prognostizierten Überlebenszeit mit der tatsächlichen verglichen wird. Mit Hilfe von Data Mining wurde bereits versucht (siehe [SG11]) solche Überlebenszeitprognosen zu erstellen. Es ergibt sich daher die interessante Fragestellung, ob mit Hilfe der regionalen Daten, ein ähnlich gutes Ergebnis wie in der oben zitierten Studie erzielt werden kann.

1.3. Problematik und Motivation

Im Wesentlichen lässt sich der Erfolg der Behandlung einer Tumorerkrankung über die Überlebenszeit des Patienten ab dem Zeitpunkt der Diagnose messen. Diese Überlebenszeit ist von vielen Faktoren abhängig und mit aktuellen Mitteln zu Beginn der Behandlung nur schwer zu prognostizieren.
Für die Behandlung des Patienten wäre eine Abschätzung der Überlebenszeit ein geeignetes Hilfsmittel für den Arzt um den Patienten optimal behandeln zu können. Da diese Zeit erst am Ende der Behandlung feststehen kann, muss eine Abschätzung anhand des initialen Zustandes des Patienten erfolgen. Diese Abschätzungen sind oftmals recht einfach gestrickt (siehe zum Beispiel [Pap06]). So werden verschiedene Risikokriterien festgelegt; je nachdem wie viele dieser Kriterien der Patient erfüllt, wird er in Risikoklassen eingeteilt, denen jeweils mittlere Überlebenszeiten zugeordnet sind. Diese Abschätzungen berücksichtigen nur wenige Parameter und sind dabei häufig stark an das Staging (siehe dazu 2.5.3) angelehnt. Zudem ist dafür die aktive Einschätzung eines Arztes notwendig; in dieser Zeit kann der Arzt nicht seiner wesentlichen Tätigkeit, der aktiven Behandlung von Patienten, nachgehen.
Mit Hilfe von Data Mining–Verfahren besteht die Hoffnung durch Berücksichtigung größerer Attributmengen und die Betrachtung langfristig angelegter und dadurch großer Datenmengen aussagekräftige Zusammenhänge entdecken zu können. Möglicherweise können daher durch das Data Mining vorher nicht entdeckte Zusammenhänge gefunden wurden.
Durch eine automatisierte Überlebenszeitprognose wäre es möglich, den Arzt von dieser Aufgabe zu entbinden und dadurch die Abschätzung zu vereinheitlichen und noch weniger subjektiv zu gestalten. Daraus würde eine bessere Vergleichbarkeit der Behandlung von Tumorpatienten über verschiedene Kliniken hinweg resultieren. Zudem könnte bei konsequenter Anwendung ärztliches Personal gezielt entlastet werden.

1.4. Zielsetzung

Zusammengefasst versucht diese Arbeit, explorativ wenige Methoden des Data Minings auf den Datensätzen des Tumorzentrums des Gesundbrunnens anzuwenden, um Aussagen über die Dauer der Behandlung einer Tumorerkrankung zu Beginn der Behandlung treffen zu können. Dabei muss klar sein, dass eine (tages–) genaue Vorhersage der Überlebenszeit nicht möglich sein kann, da diese, aufgrund individuell unterschiedlicher Patienten, Wahrscheinlichkeitsverteilungen unterliegen. Es fließen neben den dokumentierten und berücksichtigten Parametern zudem viele weitere Parameter in den Behandlungsverlauf des Patienten mit ein, die nicht alle berücksichtigt und erfasst werden können. So ist zum Beispiel die persönliche und emotionale Situation des Patienten im Tumorregister nicht dokumentiert. Beides sind Faktoren die in der Behandlung des Patienten eine wichtige Rolle spielen können. Des Weiteren finden sich neben allen sich direkt auf den Behandlungsverlauf auswirkenden Faktoren noch weitere Umwelteinflüsse, die die Überlebenszeit verändern können, wie zum Beispiel Unfälle und Ähnliches, welche eine genaue Vorhersage unmöglich machen. Einem Patienten sollte daher kein mittlerer Überlebenswert, sondern eine Verteilung der Überlebenszeit–Wahrscheinlichkeiten geliefert werden. Diese Verteilung kann anhand vorheriger ähnlicher Patientenkohorten (wie beim K–Means Clustering) abgeschätzt, oder für jeden Patienten individuell (wie beim navien Bayes Klassifikator) berechnet werden.

2. Grundlagen

2.1. Data Mining

Data Mining ist der Überbegriff für verschiedene informationstechnische und informationsverarbeitende Verfahren. Für den englischen Begriff „Data Mining“ gibt es im deutschen keine adäquate Übersetzung, die der englischen Bezeichnung in nichts nachsteht. Am nächsten kommt dem noch der Begriff der „Daten Schürfung“, also dem tiefer gehenden Suchen innerhalb eines „Datenberges“. Vereinfacht kann gesagt werden, dass Data Mining–Verfahren darauf abzielen, innerhalb einer großen Datenmenge Zusammenhänge zu erkennen, die auf den ersten Blick, oder mit einfacheren Mitteln, nicht zu erkennen wären.
Wichtig ist in diesem Kontext auch die Datenvorverarbeitung. Dies betrifft die Schritte vor dem Anwenden des eigentlichen Data Mining Verfahrens. Dabei werden die Daten so aufbereitet, dass sie automatisch verarbeitet werden können. Dazu wird beispielsweise die Formatierung der Daten geändert oder fehlende Werte behandelt. Die Datenvorbereitung nimmt häufig den größten Teil der Zeit in Anspruch, da sie nur in geringem Maße automatisiert erfolgen kann, während zum Data Mining häufig schon vorgefertigte Werkzeuge und bereits implementierte Verfahren verwendet werden können (nach [Nor12]).
Die einzelnen Data Mining–Verfahren sind für unterschiedliche Anwendungszwecke entwickelt worden. Nicht alle Algorithmen sind im gleichem Maße für eine Überlebenszeitprognose geeignet. Es können nicht beliebige Verfahren „blind“ auf die Datensätze angewandt werden, in der Hoffnung signifikante Ergebnisse zu erzielen. Deswegen musste in erster Iteration ein Überblick über einige wichtige Data Mining-Verfahren gewonnen werden. Diese werden dann anhand ihrer Eignung für die Fragestellung bewertet und genutzt. Hier dargestellt sind mehrere Verfahren, wobei nur wenige auch angewendet wurden (eine Auswahl der Algorithmen ist in 3.1 beschrieben).

2.1.1. Cross–Industry Standard Process for Data Mining

Wie ein Data Mining-Prozess prinzipiell abzulaufen hat, wird im Cross–Industry Standard Process for Data Mining (CRISP-DM) (nach [OM09]) ausführlich beschrieben und standardisiert. Der CRISP-DM Prozess ist in sechs Phasen eingeteilt, die iterativ durchlaufen werden (jede einzelne Phase ist wiederum in mehrere Schritte unterteilt, diese werden hier jedoch nicht beschrieben):

Business Understanding In der ersten Phase wird das Umfeld, in dem die Daten erfasst wurden, verstanden. Beispielsweise bedeutet dies für diese Arbeit ein grundlegendes Verständnis von Tumorerkrankungen und deren Verlauf sowie Behandlung zu erwerben. Für diese Arbeit ist dies das Einlesen in die medizinischen Grundlagen und die Einarbeitung in Data Mining–Werkzeuge.

Data Understanding In dieser Phase werden die Datensätze betrachtet und hinsichtlich ihrer Qualität bewertet. Es werden Beispieldatensätze und Metadaten betrachtet, um ein genaueres Verständnis vom Aufbau und der Erhebung der Daten zu gewinnen.

Data Preparation In dieser, in vielen Fällen sehr zeitaufwändige Phase, wird die Datenvorverarbeitung (englisch: preprocessing) vorgenommen. Eine Vielzahl möglicher Verfahren steht dabei zur Verfügung. Hier werden nur stichwortartig einige wichtige beschrieben:

- Zusammenführung von Datenbeständen aus unterschiedlichen bzw. mehreren Datenquellen
- Bereinigung des Datenbestandes von offentsichtlich falschen Datensätzen
- Auswahl der Attribute für das spätere Data Mining
- Ziehen einer Stichprobe zur Verringerung der Rechenlast
- Umgang mit fehlenden Werten
- Formatierung der Daten
- Berechnung weiterer Attribute aus den vorhandenen

Zusätzlich können noch weitere Vorverarbeitungsschritte stattfinden. Dieser Schritt geschieht zum größten Teil manuell und ist daher, wie eingangs erwähnt, sehr zeitaufwendig.

Modelling In dieser Phase werden die anzuwendenden Data Mining-Verfahren ausgewählt und angewendet. Zudem werden Gütetests für die Qualität des Analyseergebnisses und Parameter der Algorithmen festgelegt.

Evaluation Hier wird der bisherige Prozess nochmals betrachtet, um den gewählten Weg zu bestätigen. Dabei sollte sichergegangen werden, dass nichts wichtiges übersehen worden ist und somit ein in sich schlüssiger Prozess aufgestellt wurde.

Deployment In dieser letzten Phase werden die, von den ausgewählten Data Mining–Verfahren produzierten und dokumentierten Ergebnisse genutzt, beispielsweise um Artikel zu schreiben oder Programme zu implementieren.

2.1.2. Assoziationsregeln

Assoziationsregeln leiten aus einer Kombination von eingehenden Attributen mögliche Zusammenhänge ab. Dabei werden, im Gegensatz zur Korrelation, nicht nur eindimensionale Zusammenhänge betrachtet. So werden auch Regeln gefunden, die aus einer Kombination mehrerer Attribute einen Zusammenhang zu einem oder mehreren weiteren Attribut(en) herstellen (siehe [PNT06] Seite 327ff.).
Assoziationsregeln finden unter anderem in der Warenkorbanalyse von Einkäufen Anwendung. Dabei kann das Ziel der Kaufhäuser sein, Warengruppen zu finden, die besonders häufig zusammen gekauft werden um diese Artikel dann beispielsweise nahe neben einander zu platzieren. So eine Regel könnte dann zum Beispiel lauten: „Kunden, die Brot kaufen, werden auch sehr

wahrscheinlich Aufschnitt kaufen“. Allerdings sind auch kompliziertere Regeln möglich, die wie folgt lauten könnten: „Kunden, die Holzkohle und Grillanzünder kaufen, kaufen mit hoher Wahrscheinlichkeit auch Bier und Grillfleisch“.

2.1.3. Entscheidungsbauminduktion

Bei der Entscheidungsbauminduktion (siehe [IHW11], Seite 191ff.) wird anhand der gegebenen Trainingsdaten ein gerichteter Baum aufgespannt. Von der Wurzel ausgehend bilden dann zwei oder mehr Kanten Pfade zu Knoten, welche sich abermals verzweigen können. Den Abschluss des Baumes bilden die Blätter, die später die klassifizierten Entitäten enthalten. Um eine Entität zu klassifizieren, wird beginnend vom Wurzelknoten ein Pfad zu einem Blatt gesucht. Dabei wird bei jedem Knoten auf Basis der Eigenschaften der Entität eine Entscheidung getroffen. An jedem Knoten beginnen mehrere Kanten, die mit verschiedenen Aussagen bezüglich des Attributes annotiert sind. Von diesen Aussagen trifft immer genau eine auf das zu klassifizierende Objekt zu, welches dann auf der gegebenen Kante zum nächsten Knoten „wandert“. Dieser Prozess wiederholt sich, bis das Objekt an einem Blatt angekommen ist.
Ein einfacher Entscheidungsbaum kann aussehen, wie in Grafik 1 dargestellt.

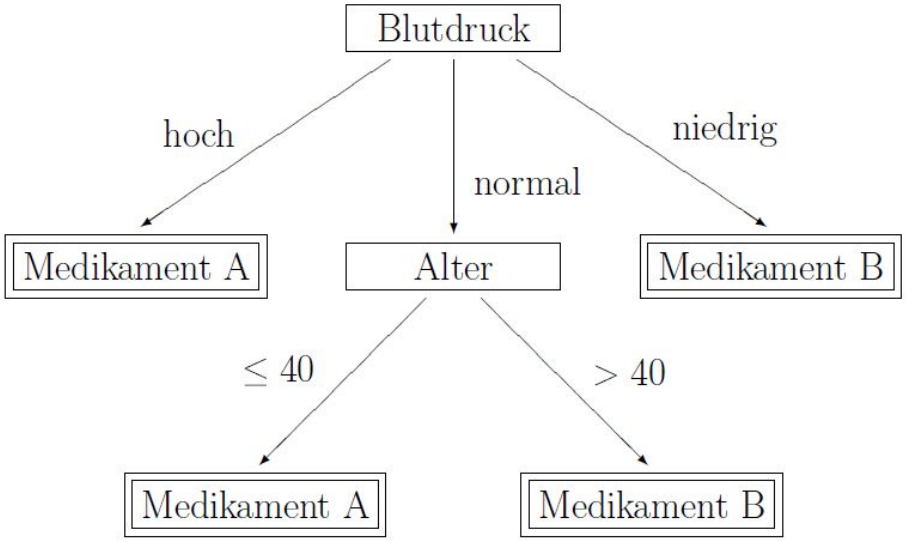

Abbildung 1: Ein einfacher Entscheidungsbaum nach [Bor10]

Dabei wird beschrieben welches fiktive Medikament einem Patienten, in Abhängigkeit von seinem Blutddruck und Alter, verabreicht werden soll.

2.1.4. K–Means Clustering

Das K–Means Clustering [Mac67] ist ein unüberwachtes Lernverfahren und versucht eine Menge von Datensätz in k Cluster, also Punktemengen mit ähnlichen Eigenschaften, zu zerlegen. Da-

für wird jeder Datensatz, beziehungsweise jede Entität, als Punkt in einem hochdimensionalen Raum betrachtet. Die Anzahl der Dimensionen entspricht dabei der Anzahl der in das Verfahren einbezogenen Attribute. Die Anzahl der Cluster k, in die die Menge aller Datensätze zerlegt werden soll, wird zu Beginn vom Anwender festgelegt. Entsprechend der gewählten Zahl k werden dann Clusterzentren möglichst weit, nach einem vorher definiertem Abstandsmaß wie zum Beispiel dem euklidische Abstand, voneinander entfernt in der Menge der Datensätze platziert. Im nächsten Schritt wird über alle Punkte des Datenbestandes iteriert und für jeden einzelnen errechnet zu welchen Clusterzentrum der Abstand bezüglich des Abstandsmaßes minimal ist. Diese Lösung ist sicherlich noch nicht optimal, also muss der Prozess wiederholt werden, um auf ein verbessertes Ergebnis zu kommen. Dazu ist es notwendig, neue Clusterzentren zu definieren, da sonst der deterministische Algorithmus die gleiche Berechnung nochmals ausführen würde. Von der geclusterten Menge der Datensätze kommt man zu neuen Clusterzentren, indem der Schwerpunkt jedes einzelnen Clusters berechnet wird. Der Schwerpunkt kann im einfachsten Ansatz als Mittelwert der zum Cluster gehörenden Punkte angesehen werden.
Diese Schritte des Berechnens neuer Clusterzentren und der Zuordnung der Punkte zu den neuen Clusterzentren wiederholen sich so lange, bis sich die Lage der Clusterzentren nicht mehr ändert, oder ein manuelles Abbruchkriterium gefunden wird (zum Beispiel: maximal 100 Durchläufe).
Die Qualität der gefundenen Cluster kann mittels der Fehlerfunktion SSE (engl. für sum of squared errors) berechnet werden. Dies ist die Summe der quadrierten euklidischen Abstände der geclusterten Datenpunkte zu ihrem jeweiligen Clusterzentrum:

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_i^j - c_j)^2$$

mit

- x_i^j zu einem Cluster j zugeordneter Datensatz,
- n_j Anzahl der zu einem Cluster j zugeordneter Datensatz,
- c_j Clusterzentren.

Dieses Fehlermaß kann auch eine Entscheidungshilfe für die Wahl der Menge an k Clustern sein. Dazu wird eine bestimme Anzahl von k Werten durchprobiert und für jedes der SSE berechnet. Diese Werte werden daraufhin in einem Diagramm verzeichnet (x-Achse Anzahl k, y-Achse Wert SSE). Diese Diagramme werden im Folgenden als k–Diagramme bezeichnet. Da mit mehr möglichen Clusterzentren der Abstand zu den Clusterzentren immer geringer wird, fällt die Kurve zu Beginn stark und nähert sich dann immer weiter 0 an. Ziel muss es sein, in der Kurve eine „Kante“ zu finden, also eine Stelle k, an der die Kurve erst steiler als erwartet fällt und danach flacher verläuft. Dieser Wert ist offensichtlich eine bessere Anzahl an k Clustern, als die umliegenden Werte, da der Abstand erst unerwartet stark sinkt, und danach die erwartete Verringerung ausbleibt, oder schwächer ausfällt.

2.1.5. Naiver Bayes Klassifikator

Mit Hilfe der naiven Bayesklassifikation (siehe [PNT06], Seite 227ff.) kann die Wahrscheinlichkeit der Zugehörigkeit von Entitäten zu bestimmten diskreten oder diskretisierten Klassen bestimmt werden. Diese Wahrscheinlichkeit wird unter der Annahme berechnet, dass die Attribute untereinander jeweils statistisch unabhängig sind. Diese vereinfacht das Klassifizierungsproblem erheblich. Bei einer Menge von Datensätzen aus realen Daten ist diese Unabhängikeitsannahme fast immer verletzt, allerdings liefert der naive Bayes Klassifikator in der Praxis trotzdem gute Ergebnisse.
Die Anwendung des naiven Bayes–Klassifikators teilt sich, wie bei allen überwachten Lernverfahren, in zwei Phasen auf. Zum einen in eine Trainingsphase, in der der Klassifikator mit validen Trainingsdaten auf das aktuelle Problem trainiert wird. Dabei ist darauf zu achten, dass beim Training keine bestimmte Attribut–Kombination durch Redundanz übertrainiert wird. Zum anderen kann nach diesem überwachten Lernvorgang die eigentliche Klassifikation erfolgen, bei der den zu klassifizierenden Entitäten aufgrund der Ausprägungen ihrer Attribute die wahrscheinlichste Klasse zugeordnet wird.
Die Klassifikation basiert auf dem Bayesschen Theorem, nach der die Wahrscheinlichkeit, dass Ereignis A eintritt, wenn Ereignis B eingetreten ist (Aposteriori-Wahrscheinlichkeit $P(A|B)$), sich auch durch die Wahrscheinlichkeit, dass B eintritt, wenn A eingetreten ist, darstellen lässt, sofern die Eintrittswahrscheinlichkeiten für A und B (Apriori-Wahrscheinlichkeiten $P(A)$ und $P(B)$) bekannt sind.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Eine Entität, repräsentiert durch n Attribute, wird danach jener Klasse zugeordnet, zu der es mit der höchsten Wahrscheinlichkeit gehört. Dazu wird für jede Entität das Maximum der folgenden Formel gesucht:

$$P(Y|X) = \frac{P(Y) \cdot \prod_{i=1}^{n} P(X_i|Y)}{P(X)}$$

mit

- X als Entität, beschrieben durch n Attribute,
- Y als Zielattribut.

Da $P(X)$ für jedes $P(Y)$ fest ist, kann das Problem auf die Maximierung des Termes $P(Y) \cdot \prod_{i=1}^{d} P(X_i|Y)$ vereinfacht werden. Die für die Berechnung notwendigen, bedingten Aposteriori– und Apriori–Wahrscheinlichkeiten werden aus den Auftrittshäufigkeiten von X und Y in den Trainingsdaten geschätzt. Für eine genaue Schätzung ist daher eine große Menge an Datensätzen unabdingbar.

2.2. Merkmalsauswahl (Feature Selection)

Die beschriebenen Data Mining–Verfahren können nur funktionieren, wenn in den Daten auch Zusammenhänge erkannt werden können. Damit mögliche Ergebnisse nicht in einem Grundrauschen der Daten untergehen oder die gewählten Data Mining-Verfahren durch ein Übermaß an Attributen „überlastet“ werden ist es notwendig, sich auf einige wenige Attribute zu beschränken. Diese Auswahl geschieht in mehreren Schritten. Zum einen werden während der Datenvorverarbeitung in einer manuellen Vorselektion Attribute aus dem Datenbestand entfernt, die keine Aussagekraft besitzen (siehe [PNT06], Seite 52ff.).
Die verbleibenden Attribute werden dann abermals selektiert. Dies kann mit Hilfe verschiedener Methoden geschehen. Einige davon werden im Folgenden (alle nach [GHJ94]) vorgestellt.

2.2.1. Naiver Ansatz

Die einfachste Art der Merkmalsauswahl ist, keine Merkmalsauswahl durchzuführen und alle Attribute so zu behandeln, als wären sie gleich wichtig und damit auch gleich gut im Bezug auf die Fähigkeit — in diesem konkreten Fall — die Überlebenszeit zu prognostizieren. Dieser naive Ansatz kann zur Validierung der anderen Merkmalsauswahlverfahren verwendet werden.

2.2.2. Backward Elimination

Die Backward Elimination geht zu Beginn von der kompletten Menge an Attributen aus. Es wird dann iterativ jedes Attribut weggelassen und dabei kontrolliert, ob sich das Ergebnis des Data Mining–Verfahrens verschlechtert hat. Wenn dies der Fall ist, wird das Attribut wieder zu der Menge der ins Data Mining mit einbezogenen Attribute hinzugenommen. Falls dem nicht so ist, wird das Attribut weggelassen. Dieser Vorgang wiederholt sich so lange, bis jedes weitere Weglassen eines zusätzlichen Attributes, schlechtere Ergebnisse liefern würde, oder bis ein manuelles Abbruchkriterium erreicht wird. Ein solches manuelles Abbruchkriterium kann beispielsweise das Erreichen einer Mindestanzahl von Attributen sein.

2.2.3. Forward Selection

Die Forward Selection ähnelt der Backward Elimination, nur beginnt das Verfahren mit einer leeren Attributmenge. Dieser werden iterativ einzelne Attribute hinzugegeben und es wird kontrolliert, ob sich das Ergebnis des Data Mining–Verfahrens verbessert. Falls dies der Fall ist, wird das Attribut zu der Menge der ins Data Mining mit einbezogenen Attribute hinzugenommen, ansonsten wird das Attribut weggelassen. Dieser Vorgang wiederholt sich so lange, bis jedes weitere Hinzunehmen eines zusätzlichen Attributes schlechtere Ergebnisse liefern würde, oder bis ein manuelles Abbruchkriterium erreicht wird.

2.2.4. Informationsgewinn (Information Gain)

Information Gain berechnet für jedes Attribut ein „Gewicht“, welches die Relevanz des Attributes für den Data Mining–Prozess darstellt. Je höher das Gewicht eines einzelnen Attributes ist, desto höher und wichtiger ist das Attribut. Die Wichtigkeit eines Attributes wird dabei über den Informationsgehalt definiert. Der Informationsgehalt ist umso höher, je größer die informationstheoretische Entropie des Attributes ist. Die Entropie ist größer je stärker die Verteilung der möglichen Ausprägungen der Gleichverteilung ähnelt. Zudem steigt die Entropie mit der Anzahl der Ausprägungen der einzelnen Attribute. Daraus resultiert der Nachteil, dass das ansonsten sehr zuverlässige Verfahren Attribute mit vielen Ausprägungen bevorzugt.

2.2.5. Expertenselektion

Neben den automatischen Merkmalsauswahlverfahren kann trotzdem eine manuelle Selektion der Attribute erfolgen. Mit der Auswahl sollte dabei eine Person betraut werden, die Erfahrung in dem zu behandelnden Gebiet mitbringt und somit fundierte Entscheidungen treffen kann, welche Attribute berücksichtigt werden sollten und welche nicht.

2.3. Datensätze

Die Datensätze wurden vom Tumorzentrum des Klinikums am Gesundbrunnen Heilbronn aufgezeichnet und archiviert. Dort wurde bereits seit Mitte der 80er Jahre ein lokales Tumorregister aufgebaut. Dieses wurde bisher nie einzeln und gezielt zu Forschungszwecken untersucht. Da das Klinikum am Gesundbrunnens alle Ausprägungen und Lokalisationen von Tumorerkrankungen behandelt, handelt es sich um einen sehr vielfältigen, großen und unübersichtlichen Datenbestand. Um diese Diversität einzuschränken und dadurch aussagekräftigere Ergebnisse zu erlangen, wurden im Hinblick auf Bitte des Darmzentrums Gesundbrunnens nur kolorektale Tumorerkrankungen berücksichtigt (ICD-10-Codes C18, C19 und C20).

Unter diesen Voraussetzungen wurden 3759 passende Datensätze im Tumordokumentationssystem der Onkologie des Gesundbrunnens Heilbronn gefunden. Die Datensätze wurden jeweils unter Berücksichtigung von bis zu 195 Attributen dokumentiert. Dabei ist zu beachten, dass nicht alle Attribute für das Data Mining verwendet werden konnten (siehe 3.3).

2.3.1. Diskretisierung der Überlebensdauer

Da die einfache, in Rapid Minder implementierte und genutzte, Version des naiven Bayes Klassifikators nicht mit stetigen Zielattributen (in diesem Fall der Überlebenszeit) umgehen kann, muss das Zielattribut diskretisiert werden. Diskretisierung bedeutet, den stetigen Wertebereich eines Attributes in einen diskreten umzuwandeln. Dazu wird der stetige Wertebereich auf mehrere diskrete „Behälter“, also Diskretisierungsintervalle, abgebildet. Für diese Abbildung gibt es ebenfalls mehrere Ansätze. Im Folgenden sind nur zwei der möglichen Ansätze vorgestellt:

Äquifrequente Diskretisierung Bei diesem Ansatz wird versucht in jedem Intervall ungefähr die gleiche Anzahl an Datensätzen zu erhalten. Daraus resultiert, dass die zeitlichen Intervalle für die einzelnen „Behälter“ unterschiedlich groß sein können.

Äquidistante Diskretisierung Bei diesem Ansatz werden die zeitlichen Intervalle, die jeweils in einen „Behälter“ abbilden, gleich groß gewählt. Bei inhomogenen Verteilungen kann sich daher die Anzahl an Attributen je „Behälter“ stark unterscheiden.

2.4. Werkzeuge

Im Rahmen der Arbeit wurde im Wesentlichen das Data Mining Werkzeug „Rapid Miner“ genutzt. Dieses und alle weiteren im Zusammenhang mit der Thesis verwendeten Werkzeuge werden im Folgenden dargestellt und ihre Aufgabe erläutert.

2.4.1. Rapid Miner

Rapid Miner [IM13] bietet die Möglichkeit, Data Mining Prozesse anhand eines intuitiv verständlichen Kontrollflusses schnell zu modellieren. In Rapid Miner sind viele Data Mining–Algorithmen implementiert. Diese können via „Drag and Drop“ in den Kontrollfluss gezogen werden. Am Ende können Ergebnistabellen und Diagramme produziert werden, wie zum Beispiel die Verteilungen jedes einzelnen Attributes, dargestellt als Histogramm.
Rapid Miner bietet diverse Möglichkeiten, Daten zu importieren. Dies kann über einen einfachen Import aus einer csv–Datei oder über eine Schnittstelle zu einer Datenbank geschehen. Ebenso können die Daten auf vielen verschiedenen Wegen wieder aus Rapid Miner exportiert werden.
Die Datenvorverarbeitung wird von Rapid Miner durch diverse vorprogrammierte Algorithmen unterstützt. Alle wichtigen Data Mining–Verfahren sind ebenso vorprogrammiert und ermöglichen daher einem Benutzer das schnelle Experimentieren mit diversen Verfahren, ohne alle selbst implementieren zu müssen. Dabei können meist jeweils wichtige Parameter des Algorithmus manuell gesetzt werden, falls der Benutzer von den Standartwerten abweichen will.
Auch stehen schon einige Möglichkeiten zur Beurteilung und Visualisierung von Ergebnissen zur Verfügung, so dass die Auswertung der Analyse ebenfalls beschleunigt wird.
Neben Rapid Miner gibt es weitere Data Mining Werkzeuge (zum Beispiel knime [Ber13] und R [r13]). Aufgrund des Funktionsumfangs und der ausführlichen Dokumentation, sowie der intuitiven Verständlichkeit wurde die Entscheidung getroffen mit Rapid Miner zu arbeiten.

2.4.2. Gießener Tumordokumentationssystem(GTDS)

Gießener Tumordokumentationssystem (GTDS) ist das Tumordokumentationssystem des Tumorzentrums Heilbronn–Franken. Das Programm hatte zum Zeitpunkt der Ausarbeitung der Thesis keine Möglichkeit einer direkten Pseudonymisierung oder Anonymisierung der dokumentierten Fälle zur wissenschaftlichen Verwendung geboten. Daher wurde ein manueller Export der Auswertungstabelle aus der Datenbank vorgenommen. Bei diesem Export (siehe Anhang A,

Tabelle 17) wurden alle Attribute, die eine direkte Rückverfolgung ermöglichen würden, nicht exportiert. Dies betrifft „NAME“, „HAUSARZT“, „VORNAME“. Ebenso wurde das Attribut „DATUM_DER_AUSWERTUNG“ nicht exportiert, da darin nur das Datum des Exportes erfasst wird, das Attribut also keinerlei Informationsgehalt hat. Der Export wurde am 13. November 2013 vorgenommen und in einer Comma Separated Value–Datei gespeichert.
Wie eingangs erwähnt, war zum Zeitpunkt der Arbeit an der Thesis keine direkte Schnittstelle zu GTDS möglich, daher sind alle Änderungen an den Datensätzen, die nach dem 13. November 2013 vorgenommen wurden, in den aktuellen Ergebnissen nicht mehr berücksichtigt.

2.4.3. Sonstige

Weitere Werkzeuge fanden zur schnellen Visualisierung von Ergebnissen und zum Schreiben der Thesis ihre Anwendung, werden hier aber nicht genauer vorgestellt. Dazu zählen unter anderem:

- Kaplan–Meier–Überlebensstatistik Excel–Makro des Universitätsklinikums Halle (Saale) für die Berechnungn und Visualisierung von Kaplan–Meier–Kurven [Sek13]
- OpenOffice und Microsoft Office für schnelle Diagrammerstellung
- Eclipse Indigo zum Programmieren des später vorgestellten (siehe Abschnitt 3.2.1) Programmes [ecl13]
- NClass zum Zeichnen des Klassendiagrammes [Tih14]

2.5. Medizinische Grundlagen

2.5.1. Tumore

Tumore sind Wucherungen innerhalb des menschlichen Körpers und können je nach Ausprägung und Gestalt von benigner oder maligner Natur sein. Benigne Tumore zeichnen sich durch ein nicht–invasives und –verdrängendes Wachstum aus. Diese sind daher für Menschen meist ungefährlich. Maligne Tumore können durch ihr invasives und schnelles Wachstum benachbartes Gewebe beeinträchtigen und zerstören, zudem können diese metastasieren und somit ihre degenerative Wirkung verstärken. Dies kann zu funktionellen Beeinträchtigung des betroffenen Organs bis zum Tod des betroffenen Gewebes führen.
Da die Behandlung eines Tumors meist weitreichende Konsequenzen für das Leben des Patienten hat, muss jede Behandlung ausführlich dokumentiert werden. Dadurch wird die Nachvollziehbarkeit der Behandlung gewährleistet. Auch kann nur so eine Abrechnung der Behandlung mit den Krankenkassen erfolgen. Zudem werden die Behandlungen auch in regionalen, nationalen und internationalen Krebsregistern erfasst. Dort werden die Datensätze dann für das gesamte Einzugsgebiet mit den gleichen Attributen dokumentiert, was bei einer rein institutionellen Dokumentation nicht gewährleistet ist. Diese größeren Register ermöglichen statistisch relevante Rückschlüsse (nach [NM99], Seite 72ff.).

2.5.2. Kaplan–Meier–Schätzer

Um die Überlebenszeiten einer Patientenkohorte in geeigneter Weise kompakt darzustellen, bedarf es Hilfsmitteln, die an die besonderen Bedingungen der Überlebenszeitanalyse angepasst sind: So ist für eine Betrachtung der Letalität einer bestimmten Erkrankung nur die Menge der Patienten interessant, die an eben dieser Krankheit gestorben sind (und damit auch nur deren Überlebenszeiten). In einer erhobenen Stichprobe von Patienten sind allerdings mehr Patienten vorhanden, als jene, auf die diese Definition zutreffen. So gibt es Patienten, die aus anderem Grund während des Behandlungszeitraumes sterben (sogenannte „drop–outs“, zum Beispiel durch Unfälle etc.) oder Patienten, die die Behandlung während ihres Verlaufes selbstständig abbrechen und keine weitere Nachverfolgung wünschen (als „lost to follow–up“ bezeichnet). Zu Beginn der Behandlung ist allerdings nicht absehbar, wie sich der Patient in Zukunft verhalten wird, daher kann eine Zensur nicht schon zu Beginn stattfinden. Würden zensierte Patienten im Nachhinein nicht mehr in Betracht gezogen, würde die Information, dass der Patient mindestens bis zum Zeitpunkt der Zensur überlebt hat, verloren gehen.
Um all diese Rahmenbedingungen zu berücksichtigen und trotzdem eine übersichtliche Form der Darstellung zu ermöglichen, wurde der Kaplan–Meier–Schätzer (oder auch die Kaplan–Meier–Kurve bzw. Kaplan–Meier–Diagramm) entwickelt. Dieser trägt in einem zweidimensionalen Koordinatensystem auf der x–Achse die fortschreitende Zeit und der y–Achse den Anteil der noch lebenden Patienten auf. Dabei werden jedoch auch Zensuren, also Patienten, die aus einem beliebigen Grund (zum Beispiel Unfall oder Verweigerung der Weiterbehandlung) aus der Behandlung ausscheiden, berücksichtigt. Zensuren werden zu dem Zeitpunkt, an dem sie eintreten mit einem Kreuz auf der Kaplan–Meier–Kurve gekennzeichnet. Die Höhe der Kurve ändert sich durch eine Zensur nicht. Tumorbedingte Todesfälle ändern allerdings die Höhe der Kurve, welche durch folgende Formel für jeden Zeitpunkt t_i (mit $i \in \mathbb{N}$ und $i \leq$ Anzahl Patienten), an dem ein Ereignis (Tod oder Zensur eines Patienten) eintritt bestimmt wird:

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i} = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

mit:

- $\hat{S}(0) = 1$,
- d_i Patienten, bei denen der tumorbedingte Tod zum Zeitpunkt t_i eingetreten ist,
- n_i Anzahl der Patienten, die zum Zeitpunkt t_i unter Risiko.

Wobei „unter Risiko“ bedeutet, dass der Patient aktuell noch lebt und für diesen Patienten noch keine Zensur eingetreten ist.

Die Kaplan–Meier–Kurve beginnt also zum Zeitpunkt 0 bei 100% und sinkt dann monoton, bis entweder der letzte Patient verstirbt (Kaplan–Meier–Kurve sinkt auf 0%) oder eine Zensur eintritt (Ende der Kaplan–Meier–Kurve auf der Höhe, die sie vor der Zensur hatte). Zu jedem anderen Zeitpunkt $t_i \neq (t_0, t_{Ende})$ stellt die Höhe der Kurve das Produkt der Wahrscheinlich-

keiten des Überlebens aller vorherigen Zeitpunkte $t_{i-1}, t_{i-2}, ...$ dar. Da jede Wahrscheinlichkeit aus $\mathbb{R}$ und in $[0, 1]$ ist, ist damit auch das monoton fallende Verhalten der Kurve erklärt (nach [Efr88]).

2.5.3. Staging

Staging bezeichnet die Einteilung eines Tumors in verschiedene Stadien. Am verbreitetsten ist dabei das Tumor Nodes Metastasis (TNM)-Staging. Dieses Verfahren beruht auf einer hauptsächlich dreiachsigen Klassifizierung [tnm14] und ist abhängig von der lokalisation des Tumors. Hier ist das TNM-Staging für kolorektale Tumore angegeben.

T (Tumor) Der T–Wert beschreibt die räumliche Ausdehnung des Primärtumors. Er wird in Abstufungen von T0 (keine Infiltration) bis T4 (Infiltration von Nachbarorganen oder des Bauchfells) eingeteilt.

N (Lymphknoten (engl. nodes)) Der N–Wert beschreibt, ob der Tumor schon Lymphknotenmetastasen ausgebildet hat. Dieser N-Wert wird stufenweise von N0 (keine Anzeichen für Lymphknotenbefall) bis N1,2 oder 3 (steigender Lyphknotenbefall) angegeben.

M (Metastasen) Der M–Wert beschreibt beschreibt, ob der Tumor bereits Fernmetastasen gebildet hat, und wird mit M0 (keine Anzeichen für Fernmetastasen) und M1(Fernmetastasen vorhanden) angegeben

Zusätzlich dazu sind unter anderem folgende weitere Staging-Kategorien optional möglich:

L (Lymphgefäße) beschreibt, ob der Tumor bereits in Lymphgefäße eingedrungen ist (L1), oder ob dies nicht der Fall ist (L0)

V (Venen) beschreibt, ob der Tumor bereits in Venen eingedrungen ist (L1), oder ob dies nicht der Fall ist (L0)

2.5.4. Grading

Das Grading [gra14] ist, trotz ähnlichem Namen, nicht zum Verwechseln mit dem Staging und dient der Bestimmung von Gewebeeigenschaften des Tumors, also wie gut der Tumor noch von gesundem Gewebe zu unterscheiden ist. Das Grading wird in der Pathologie durch eine mikroskopische Untersuchung angefertigt. Im Zuge dieser pathologischen Untersuchung wird der Tumor in eine der folgenden Kategorien eingeteilt:

Grad 1 (G1): gut differenziertes bösartiges Gewebe („low-grade“), hohe Übereinstimmung mit Ursprungsgewebe

Grad 2 (G2): mäßig differenziertes bösartiges Gewebe

Grad 3 (G3): schlecht/niedrig differenziertes bösartiges Gewebe

Grad 4 (G4): nicht differenziertes bösartiges Gewebe (undifferenziert bzw. anaplastisch) („high-grade“)

Grad 9 (G9): Grad der Differenzierung nicht zu beurteilen

Patienten mit einem niedrigeren Grading haben in der Regel eine bessere Prognose auf den Krankheitsverlauf und höhere Heilungschancen.

3. Entwurf

3.1. Auswahl möglicher Verfahren

In Abschnitt 2.1 sind mehrere Data Mining Verfahren beschrieben, die alle potentiell geeignet waren, aussagekräftige Ergebnisse für die gegebene Aufgabenstellung zu produzieren. Da der Umfang einer Bachelorarbeit es nicht zulässt, alle Verfahren mit der gleichen Tiefe und Gründlichkeit zu behandeln, musste entschieden werden, welche Verfahren getestet werden.
Es wurden, der Einfachheit halber weit verbreitete, bekannte und einfache zu verstehende Verfahren gewählt (siehe [al.07]). Zum einen fiel dabei die Entscheidung auf das K–Means Clustering, da es die Möglichkeit versprach, für ganze Patientengruppen Überlebenszeitprognosen zu liefern. Zudem wurde mit dem naiven Bayes Klassifikator ein ebenfalls weit verbreitetes Verfahren gewählt. Die Auswahl dieser beiden Verfahren ist besonders interessant, da zwei komplett unterschiedlich Ansätze gewählt werden. Beim K–Means Clustering werden die Eigenschaften von Patientengruppen betrachtet und erst im letzten Schritt ihre Bedeutung für die Überlebenszeit in die Überlegung mit einbezogen. Beim naiven Bayes wird im Gegensatz dazu für jeden Patient eine Vorhersage auf die Überlebenszeit errechnet. Die Verfahren wurden bereits in Kapitel 2.1 beschrieben. Im Folgenden wird nur noch die Anwendung auf die konkrete Fragestellung erläutert.

3.1.1. K–Means Clustering

Auf den ersten Blick ist nicht ersichtlich, wie ein Clustering zu einer Überlebenszeitvorhersage genutzt werden kann. Eine direkte Vorhersage ist mit dem K–Means Clustering auch nicht möglich, weshalb folgender Umweg gewählt wurde: Zuerst wurde ein Clustering auf dem Datenexport aus dem Tumordokumentationssystem durchgeführt, wobei die Überlebensdauer (und auch ob der Tumortod eintraf) nicht im Clustering berücksichtigt wurden. Somit werden auf Basis des euklidschen Abstandes als Abstandsmaß ähnliche Patientengruppen gefunden. Dieser geclusterte Menge von Datensätzen wird danach (über „PAT_ID“ als Schlüssel, siehe Anhang A, Tabelle 17) mit der Überlebensdauer und dem Tumortod–Flag wieder zusammengeführt. Dadurch wird es möglich, für jeden Cluster eine Kaplan–Meier–Kurve zu zeichnen. Durch die grafische Darstellung kann danach leicht abgelesen werden, ob sich die Kaplan–Meier–Schätzer und damit die Überlebenszeiten für die einzelnen Cluster voneinander unterscheiden.
Für alle späteren Patienten kann ihr Abstand zu den jeweiligen Clusterzentren berechnet werden. Jener Cluster, bei welchem das Abstandsmaß für den Patient minimal ist, liefert mit dem Kaplan–Meier–Schätzer eine Verteilung der Überlebenszeiten für den Patienten.

Bei der Auswahl einer geeigneten Anzahl von Clusterzentren k, muss, wie in Abschnitt 2.1.4 beschrieben, das k–Diagramm betrachtet und dabei nach Kanten in der Kurve gesucht werden. Es kann allerdings vorkommen, dass in einer Kurve mehr als eine Kante zu erkennen ist. Dann muss eine Entscheidung für eine der Kanten gefällt werden, im besten Fall für die stärkere. Die stärkere Kante lässt sich ermitteln indem ein Wert interpoliert wird, an der Punkt ungefähr liegen sollte, sofern dort keine Kante wäre:

$$z_i = \frac{x_{i-1} + x_{i+1}}{2}$$

mit

z_i als interpolierter Wert an der Kantenstelle,

x_{i-1},x_{i+1} als Nachbarpunkt der Kantenstelle.

Wobei sich die Abweichung vom eigentlichen Wert des mittleren Abstandes an der Kante als $\Delta x = z_i - x_i$ mit x_i als Wert an der Kantenstelle darstellt. Die Stelle x_i mit dem größeren Δx wird als besserer Wert für die Anzahl der Clusterzentren angesehen.

3.1.2. Naiver Bayes Klassifikator

Der naive Bayes Klassifikator kann die Wahrscheinlichkeit der Zugehörigkeit (in Prozent) einer Entität zu einer konkreten Klasse bestimmen. Es ist daher offensichtlich die Überlebensdauer zu diskretisieren. Diese Überlebenszeitintervalle sind dann das Zielattribut für die naive Bayes Klassifikation.
Nach dem Training mit den Datensätzen kann der naive Bayes Klassifikator für jeden neuen Patienten eine Wahrscheinlichkeit der Zugehörigkeit zu jedem Intervall berechnen. Man erhält also eine individuelle Vorhersage des Überlebenszeitwahrscheinlichkeitsverteilung für jeden Patienten.
In einem naiven Ansatz könnte ein zu klasifizierender Datensatz einfach auf das Mehrheitsintervall klassifiziert werden. Da der naive Bayes Klassifikator allerdings für jeden Datensatz ein Wahrscheinlichkeitshistogramm der Zugehörigkeit zu den Klassen liefert, wird im Folgenden mit dem Histogramm gearbeitet.

3.2. Merkmalsauswahl (Feature Selection)

Die Funktionsweise der verschiedenen Verfahren wurde bereits in Abschnitt 2.2 behandelt. Hier beschrieben ist jeweils nur der konkrete Anwendungsfall. Es wurde dafür ein Werkzeug entwickelt, welches die manuelle Vorselektion übernimmt, sowie alle notwendigen Vorverarbeitungsschritte durchführt.

3.2.1. Werkzeug zur Datenvorverarbeitung

Wie in Abschnitt 2.4.1 erwähnt, bietet Rapid Miner einige Möglichkeiten zur Datenvorverarbeitung. Da diese allerdings immer sehr spezifisch sind und auf die konkreten Daten angepasst werden muss, kann Rapid Miner in diesem Fall nicht alles leisten, was für eine korrekte und komplette Vorverarbeitung notwendig ist. Es wurde daher beschlossen mit Hilfe von Eclipse Indigo

eine Java Anwendung zu entwickeln, um alle Schritte der Datenvorverarbeitung abzuwickeln. Dieses Programm wurde speziell für den Export aus GTDS entwickelt.
Das Programm füllt also die Lücke zwischen dem Datenexport aus GTDS und der Verarbeitung der Daten mit Rapid Miner. Folgende Aufgaben werden durch die Anwendung übernommen:

- Berechnung des Behandlungszeitraumes,
- Behandlung von fehlenden Werten,
- Normalisierung von numerischen Attributen,
- Berechnung der Überlebenszeit,
- Berechnung des Alters,
- Entfernung redundanter oder nach anderen Kriterien überflüssiger Spalten,
- Berechnung des Union Internationale Contra le Cancer (UICC)-Wertes aus den Werten des TNM-Stagings,
- Vorbereitung der Überlebens- und Therapiezeiten für die Darstellung in einer Kaplan–Meier–Kurve.

In Folge der Bearbeitung des Exportes durch das Werkzeug werden also Werte geändert, Attribute gelöscht oder hinzugefügt und neue Werte berechnet.

Der Quellcode und die Dokumentation des Programms ist der Thesis beigelegt. Das UML–Diagramm (Diagramm 2) zeigt die Struktur des Programms. Der Übersichtlichkeit halber wurden die Methoden, in denen die Datenvorverarbeitung stattfindet, nicht modelliert. Diese Methoden waren ausschließlich in der Klasse „Data“ zu finden. Die eigentliche Datenvorverarbeitung findet in der Klasse „Data“ statt. Teilweise werden Aufrufe, die die Abbildung des TNM–Stagings auf das UICC–Staging bewirken sollen, an die Klasse „UICC“ weitergeleitet. Die Methoden zur Datenvorverarbeitung (in „Data“) werden aus der Klasse „Business“ heraus aufgerufen. Die Klasse „Business“ enthält vom „CommandParser“ Befehle und leitet diese weiter, oder arbeitet sie, soweit möglich, selbst ab. Die Klasse „CommandParser“ wird von der „Starter“-Klasse heraus aufgerufen und leitet daraufhin solange alle Befehle an die Klasse „Business“ weiter bis ein *exit*–Befehl kommt.

Der unbearbeitet Export enthält 195 verschiedene Attribute. Viele davon sind zum Data Mining nicht geeignet, da Zusammenhänge die innerhalb dieser Attribute gefunden werden würden, rein zufälliger Natur wären. Daher mussten vor der eigentlichen Datenvorverarbeitung durch eine manuelle Merkmalsauswahl die potentiell interessanten Attribute herausgefiltert werden. Ein Attribut kann sich auf mehrere Arten für die spätere weitere Verarbeitung disqualifizieren. Da in dieser Arbeit aus dem Initialstadium des Patienten eine Vorhersage auf die Überlebenszeit getroffen werden sollte, sind Attribute, die erst im Verlauf der Behandlung erfasst werden, uninteressant, dies betrifft vor allem Merkmale die den Zustand nach einer Intervention betrachten. Alle Attribute, die einen einzelnen Zeitpunkt beschreiben, sind nach der Berechnung von Über-

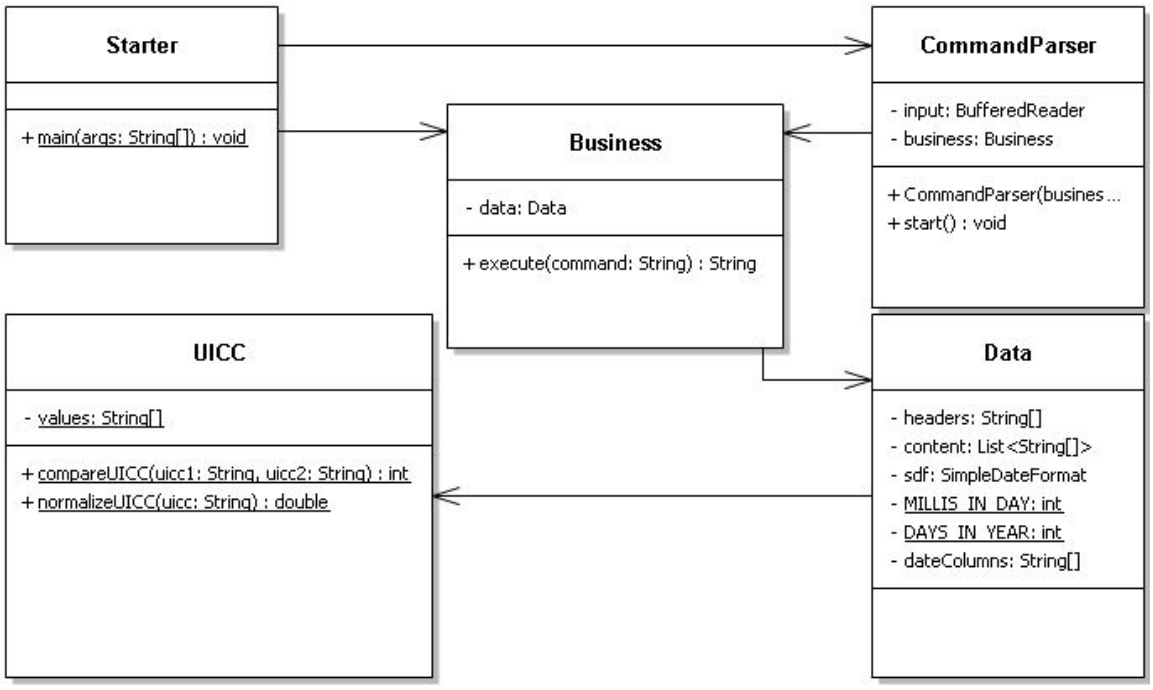

Abbildung 2: UML–Diagramm Vorverarbeitungs–Werkzeug

lebensdauer und Alter des Patienten ebenfalls nicht mehr von Interesse.
Eine weitere große Gruppe der uninteressanten Attribute bilden von GTDS generierte oder aggregierte Felder, wie IDs oder die Zusammenfassung mehrerer anderer Merkmale in einem einzigen Attribut.
Auch gibt es einige Felder, die einen echten Freitext enthalten, also nicht mit einfachen Mitteln automatisch ausgewertet werden können. Diesen Freitext zu analysieren hätte allerdings die Komplexität dieser Bachelorthesis überstiegen.
Leider gibt es auch noch eine weitere Menge an Attributen, die teilweise sogar sehr interessant für die Auswertung wären. Jedoch disqualifizieren sich alle Attribute dieser Menge für eine Auswertung dadurch, dass sie in keinen oder zu wenigen Fällen erfasst wurde, um daraus zuverlässige Aussagen gewinnen zu können. Dies betrifft allerdings nicht alle Attribute, die insgesamt betrachtet in zu wenigen Fällen genutzt wurden, da bei manchen sinnvolle Ersatzwerte für die fehlenden Werte gefunden werden können. Für die Arbeit wurde dabei die Grenze gesetzt, dass das Attribut mindestens in der Hälfte der Fälle dokumentiert worden sein musste, um im späteren Verlauf für das Data Mining in Betracht gezogen zu werden.
Ein paar wenige Attribute (die zum TNM-Staging und „LETZTE_INFO_DATENART“ sowie „TODESURSACHE“) tauchen ebenfalls in der späteren Auswertung nicht auf, da sie in der Datenvorverarbeitung durch andere, aus ihnen errechneten Werte ersetzt werden (in diesem Fall „UICC“ und „EVENTS“, mehr dazu in 3.3).
Eine Übersicht über alle Attribute und eine Kurzzusammenfassung, der Begründung, warum sie nicht für das Data Mining in Betracht gezogen wurde, ist in Tabelle 17 in Anhang A zusammengestellt.

3.2.2. Backward Elimination

Bei der Backward Elimination wurden, soweit durch Rapid Miner möglich, Parameter manuell festgelegt. Zum einen wurde bestimmt, dass so lange Attribute entfernt werden, bis jedes weitere Entfernen eine Verschlechterung der Vorhersage nach sich ziehen würde, zum anderen wurde zumindest theoretisch erlaubt, dass alle Attribute entfernt werden können. Diese Einstellungen wurden getroffen, um keines der möglichen, durch die Backward Elimination gefundenen, Attribute zu verlieren. Es wurde also keine Mindest– oder Maximalanzahl an Attributen festgelegt, welche ins Data Mining einzubeziehen sind. Gerne wäre konkrete Einstellungen vorgenommen worden, welche festgelegt hätten, dass genau n Attribute von der Backward Elimination ausgesucht werden sollten. Leider ermöglicht Rapid Miner eine solche Einstellung nicht. Daraus resultiert eine geringere Vergleichbarkeit mit anderen Merkmalsauswahlverfahren.

3.2.3. Forward Selection

Bei der Forward Elimination wurde festgelegt, dass so lange Attribute hinzugenommen werden sollten, bis aus einem weiteren Hinzunehmen keine Verbesserung der Vorhersage resultiert. Zudem wurde prinzipiell ermöglicht, dass alle Attribute in das Data Minng mit einbezogen werden könnten. Auch bei der Forward Selection stellte Rapid Miner die gleichen Grenzen wie bei der Backward Elimination, so dass auch dieses Verfahren nur bedingt mit anderen Merkmalsauswahlverfahren vergleichbar ist.

3.2.4. Manuelle Expertenselektion

Fr. Dr. Sylvia Bochum, Ärtzliche Koordinatorin des Tumorzentrums Heilbronn–Franken, übernahm freundlicherweise die manuelle Expertenselektion. Sie kam dabei auf eine Einteilung der Attribute in 5 Klassen mit ähnlicher Aussagekraft bezüglich der Überlebenszeit.
Eine sehr hohe Aussagekraft bezüglich der Überlebenszeit des Patienten besitzen laut Fr. Bochum folgende Attribute:

- Anzahl Metastasen,
- Anzahl Tumore.

Diese beiden Faktoren sollten einfach mit der Überlebenszeit korreliert sein: Je mehr Metastasen/ Tumore, desto kürzer überlebt der Patient.

Eine natürlicherweise noch hohe Bedeutung, aufgrund ihrer Eigenschaft als Staging-Faktoren besitzen laut Fr. Bochum folgende Attribute:

- UICC-Staging,
- Erste R-Klassifikation,

- Histologie Grading,
- L-Kategorie des TNM–Stagings (P_L),
- L-Kategorie des TNM–Stagings (P_V).

Die nächste Abstufung bildet der

- ICD10-Code.

Da unterschiedliche Tumorentitäten unterschiedlich gute beziehungsweise schlechte Prognosen haben (z.B hat ein Pankreaskarzinom in der Regel eine schlechtere Prognose als ein Prostatakarzinom). Eine letzte Gruppe der Attribute mit noch nicht niedriger Aussagekraft bilden die folgenden beiden Merkmale:

- Primärtherapie zur Behandlung,
- OP Intention (kurativ, pallitiv, etc.).

Alle weiteren Attribute, die nach der manuellen Vorselektion noch übrig geblieben sind, wurden von Fr. Bochum als Attribute mit niedriger Aussagekraft bezüglich der Überlebenszeit eingeteilt. Dies waren:

- Geschlecht,
- Alter,
- Behandlungsanlass,
- Erfassungsanlass,
- Arzt Anlass.

3.2.5. Information Gain

Information Gain liefert, wie in Abschnitt 2.2.4 erwähnt eine Rangfolge. Um eine Vergleichbarkeit mit der Expertenselektion herzustellen, wurde die Merkmalsauswahl, mithilfe des Information Gain–Verfahrens, jeweils viermal durchgeführt und dabei die $2, 7, 8, 10$ besten Attribute berücksichtigt. Dies entspricht quantitativ der Einteilung der Attribute in Klassen von ähnlicher Aussagekraft bezüglich der Überlebenszeit von Fr. Dr. Bochum.

3.3. Vorverarbeitung der Quellattribute

In Abschnitt 3.2.1 wurde das Werkzeug, welches für das Vorverarbeitung der Daten entwickelt wurde, in seiner Funktionalität bereits grob umrissen. Die genauen Schritte der Datenvorverarbeitung sind im Folgenden genauer erläutert. Tabellen mit den Attributen vor (Tabelle 17) und nach (Tabelle 18) der Vorverarbeitung sind im Anhang A zu finden. Tabelle 18 beschreibt zudem kurz alle Vorverarbeitungsschritte, die mit dem jeweiligen Attribut vorgenommen wurden.

Zur Berechnung des Behandlungszeitraumes (oder auch Überlebenszeit, da das Überleben des Patienten bis zu Ereignis X gemessen wird. Wobei X nicht der Tod des Patienten sein muss) muss theoretisch nur die Differenz zwischen Diagnosedatum und Behandlungsenddatum gebildet werden. Das Ende der Behandlung kann durch verschiedene Ereignisse herbeigeführt werden, zum Beispiel durch den Abbruch der Behandlung durch den Patienten oder den (unter Umständen auch tumorbedingten) Tod des Patienten. Daher muss zuerst geklärt werden, welcher Fall bei dem konkreten Patienten vorliegt und das entsprechende Datum zur Berechnung herangezogen werden. Das Diagnosedatum ist in fast allen Fällen explizit gegeben. Falls eines der beiden zur Differenzbildung notwendigen Daten nicht gegeben ist, muss ein Ersatzwert gefunden werden. Dieser ist für das Behandlungsenddatum das letzte Datum des Datensatzes. Ein geeigneter Ersatzwert für das Diagnosedatum ist das früheste Datum des Datensatzes, abgesehen vom Geburtsdatum des Patienten. Aus der Differenz dieser beiden Daten (beziehungsweise deren Ersatzwerten) kann dann die Behandlungsdauer (in Tagen) berechnet werden. Das Attribut, welches die Behandlungsdauer enthält, wurde im Datensatz „TIMES“ benannt.

Analog zur Berechnung der Behandlungsdauer wird das Alter berechnet. Das erste Datum ist dabei das Geburtsdatum des Patienten (in allen Datensätzen vorhanden). Da der Status des Patienten zu Behandlungsbeginn interessant ist, wird als zweites Datum das Diagnosedatum zu Rate gezogen. Falls dies nicht vorhanden ist, wird wie oben beschrieben ein Ersatzwert gefunden. Das berechnete Alter ist im Datensatz unter „ALTER“ zu finden.

Nachdem das Alter berechnet wurde, wurde das Attribut zudem normiert. Eine Normierung ist notwendig, um bei einer späteren Berechnung eines Abstandsmaßes (zum Beispiel beim Clustering) kein Attribut stärker zu gewichten, nur weil das Intervall aus dem Werte geschöpft werden können bei dem jeweiligen Attribut größer ist. Dies wird an folgendem Beispiel deutlich: Das Alter des Patienten nimmt in dem Datensatz Werte zwischen 0 und 97 Jahren an, während die L–Kategorie des TNM–Stagings nur Werte zwischen 0 und 2 annehmen kann. Dies würde bedeuten, dass das Alter bei der Berechnung eines euklidschen Abstandsmaßes, welches entscheidend ist für die Bildung von Clustern beim K–Means Clustering, ungefähr 50-fach stärker gewichtet wird als die L-Kategorie des TNM-Stagings. Alle Werte wurden auf ein $[0, 2]$ Intervall normiert. Die Begründung, warum auf dieses Intervall dieser Größe normiert wurde, wird in Kapitel 3.3.1 erläutert.

Es muss bei der Vorverarbeitung mit vielen fehlenden Daten gearbeitet werden. Daher werden bei den betroffenen Attributen gegebenenfalls Ersatzwerte gesucht. Bei numerischen Spalten, wie den diversen Kategorien des TNM–Stagings, werden fehlende Werte durch einen „0“-Wert ersetzt. Da häufig nur N– oder M–Werte nicht dokumentiert wurden, konnte anhand Tabelle 1 daraus geschlossen werden, dass das nicht dokumentieren dieses Wertes nur auf eine Erleichte-

UICC	TNM		
0	Tis	N0	M0
I	T1, T2		
IIA	T3		
IIB	T4a		
IIC	T4b		
III	Jedes T	N1, N2	
IIIA	T1, T2	N1a	
	T1	N2a	
IIIB	T3, T4a	N1	
	T2, T3	N2a	
	T1, T2	N2b	
IIIC	T4a	N2a	
	T3, T4b	N2b	
	T4b	N1, N2	
IVA	Jedes T	Jedes N	M1a
IVB	Jedes T	Jedes N	M1b

Tabelle 1: TNM zu UICC Mapping nach [RF11] Seite 207

rung der Arbeit der Dokumentare und nicht auf eine fehlende Dokumentation zurückzuführen ist.
Bei allen diskreten Attributen bietet GTDS eine Ausprägung des Attributes an, welche „nicht erfasst(e)“ oder „unbestimmte“ Werte zusammenfasste (häufig als „X“–Ausprägung des Attributes). Dieser Wert wurde auch bei nicht dokumentierten Feldern als Ausprägung des entsprechenden Attributes gesetzt.
Alle weiteren fehlenden Werte anderer Attribute (betrifft nur „GESCHLECHT“) wurden durch den Medianwert, in diesem Fall „M“(männlich) ersetzt.

Während ein Großteil der Vorverarbeitungs–Operationen mehr oder weniger trivial ist, sei die Berechnung des UICC-Wertes aus dem TNM-Staging hier kurz genauer dargestellt. Diese Umrechnung ist eine Abbildung, welche wie in Tabelle 1 beschrieben, vorgenommen wird. Es wird im Behandlungsverlauf ein klinisches und ein pathologisches TNM–Staging erfasst. Das klinische Staging erfolgt mit Hilfe von bildgebenden Verfahren und anderen nicht invasiven Untersuchungsmethoden. Der pathologische TNM wird nach Entnahme einer Gewebeprobe, zum Beispiel während einer OP, ermittelt. Für die Arbeit ist prinzipiell der zuerst erfasste Wert interessant, da der Initalzustand des Patienten relevant ist. Folglich wurde immer der zuerst erfasst und dokumentierte Wert für die Abbildung herangezogen. Falls beide Werte am gleichen Tag dokumentiert wurden, wurde der pathologische TNM für die Abbildung verwendet, da er standardisierter erfasst wird und weniger subjektiv ist.
Da innerhalb des UICC-Stagings eine natürliche Ordnung vorliegt, wird nach der Ermittlung des UICC-Wertes dieser ebenfalls auf ein $[0, 2]$-Intervall normiert. Die Normierung erfolgt, indem jeder UICC-Staging Stufe eine fortlaufende Zahl zugeteilt wird, beginnend bei 0 für UICC 0 bis zu 10 für UICC IVB. Anhand dieser Werte wird anschließend die Normierung vorgenommen.

3.3.1. Merkmalsdiskretisierung

Neben den numerischen Attributen und den Staging-Attributen gibt es Attribute, die sich nicht in eine natürliche Reihenfolge bringen lassen. Dazu zählt zum Beispiel das Attribut „ARZT_ANLASS“. Da keine natürliche Ordnung der Attribute gefunden werden kann, musste dafür ein anderes Abstandsmaß (für das K–Means Clustering) gesucht werden. Attribute dieser Art werden daher in mehrere Spalten aufgeteilt, eine pro möglicher Ausprägung des Attributes. Diese neuen Attribute enthalten ausschließlich Binärwerte. Dieser Wert ist immer nur für eines dieser neu entstandenen Attribute „true“, genau dann, wenn das neue Attribut die ursprüngliche Ausprägung des Attributes beschreibt.
Für das Abstandsmaß bedeutet dies, dass zwei Datensätze, die sich nur in einem so behandelten Attribut unterscheiden, den Abstand 2 haben, da sie sich durch die Aufteilung in mehrere Attribute in genau 2 Attributen unterscheiden. Um so behandelte Attribute nicht stärker zu gewichten als andere Attribute, wurden numerische Attribute auf ein $[0, 2]$-Intervall normiert.

3.4. Ansätze zur Diskretisierung der Überlebensdauer

Da zu Beginn des Verfahrens keine sinnvolle Anzahl an Diskretisierungsintervallen vorgegeben werden konnte, wurden mehrere Möglichkeiten durchprobiert. Dafür wurde die Entscheidung getroffen, die Diskretisierung für $2, 3, 4, 5, 10$ Diskretisierungsintervalle zu testen.
Da vor der Diskretisierung eine zufällige Aufteilung des gesamten Datenbestandes in eine Trainings- und eine Testdatenmenge erfolgt, sind die Verteilungen nicht bei jeder Versuchsdurchführung exakt so wie im Folgenden beschrieben. Die im Anhang A verzeichneten Verteilungen (Tabellen 22 bis 29) ähneln aber auf jeden Fall den tatsächlichen Verteilung der einzelnen Versuchsdurchführungen stark.

3.4.1. Äquifrequente Diskretisierung

Bei der äquifrequenten Diskretisierung ergaben sich, exemplarisch für zwei Diskretisierungsintervalle, die in Grafik 2 dargestellte Verteilung. Alle weiteren Verteilungen sind im Anhang A von Tabellen 22 bis Tabelle 25 zu finden.

Intervall (Überlebenszeit in Tagen)	**Anteil in %**
$[-\infty; 526, 5]$	$49, 6$
$[526, 5; \infty]$	$50, 4$

Tabelle 2: Äquifrequente Diskretisierung in 2 Intervalle

Hierbei ist die ungefähr homogene Verteilung der Datensätze zu beachten, allerdings unterscheiden sich die Intervalle sehr was die zeitliche Größe anbelangt, da die gesamten Datensätze Überlebenszeiten aus dem Intervall $[0, 27831]$ enthalten.

3.4.2. Äquidistante Diskretisierung

Da die Überlebenszeiten einen starken Ausreißer hatten, siehe Abbildung 3 (in der Abbildung nicht mehr erkennbar, ein Wert bei 27831 Tagen), wurde vor der Diskretisierung eine Ausreißeranalyse mit Elimination der Ausreißer vorgenommen.

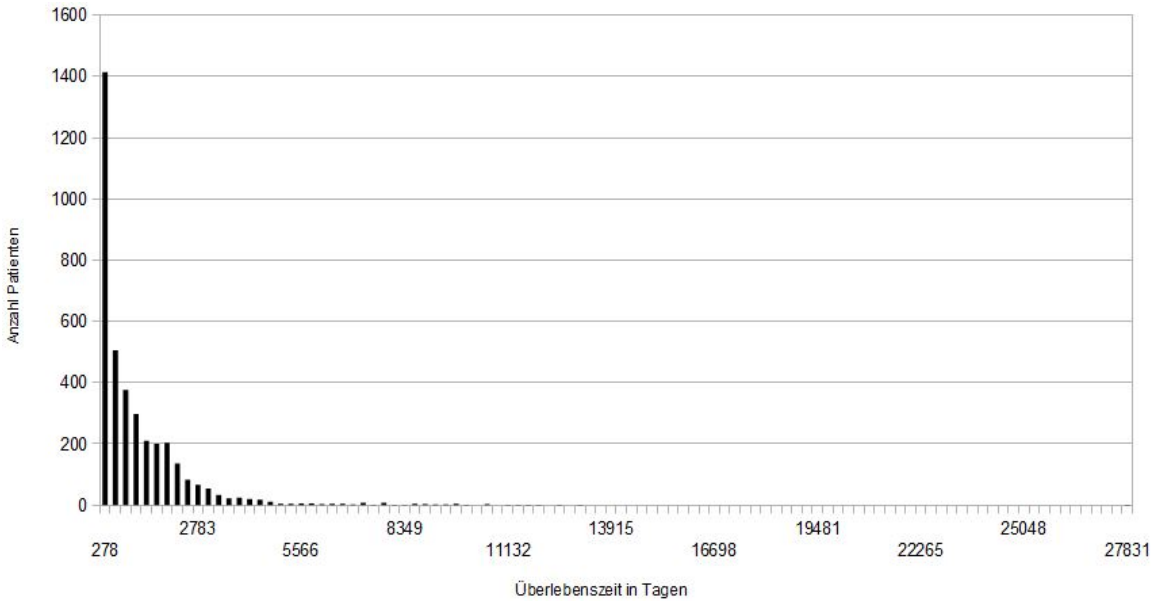

Abbildung 3: Überlebenszeiten-Histogramm

Bei der äquifrequenten Diskretisierung ergab sich dann, hier exemplarisch für zwei Diskretisierungsintervalle dargestellt, die in Abbildung 3 dargestellt Verteilung. Alle weiteren Verteilungen sind im Anhang A in den Tabellen 26bis Tablle 29 zu finden.

Intervall (Überlebenszeit in Tagen)	**Anteil in %**
$[-\infty; 6513,5]$	$99,9$
$[6513,5; \infty]$	$0,1$

Tabelle 3: Äquidistante Diskretisierung in 2 Intervalle

Trotz Ausreißerelimination ist hier schon zu erkennen, dass in dem vorderen Intervall deutlich mehr Elemente enthalten sind als in dem hinteren.

3.5. Konkrete Fehlermaße

Zur Evaluation der Güte der Ergebnisse, muss in jedem Fall nach dem eigentlichen Data Mining–Prozess ein Fehlermaß gefunden werden. Zudem muss neben der Fehlerberechnung ein Ansatz

gefunden werden, um die Qualität des vorher berechneten Data Mining–Prozesses bestimmen zu können, dies kann beispielsweise durch Berechnung von Erwartungswerten geschehen.

3.5.1. Naiver Bayes Klassifikator

Bei der Anwendung des Bayer Klassifikator gibt es prinzipiell zwei Fälle: Im ersten Fall wurde der konkrete Datensatz wurde richtig klassifiziert, im zweiten falsch. Dies ist recht einfach durch einen Vergleich des prognostizierten Überlebenszeit-Intervalls mit dem tatsächlichen Überlebenszeitintervall bei der Klassifikation innerhalb der Testdatenmenge zu realisieren. Ein erster primitiver Ansatz für ein Fehlermaß könnte also sein die Genauigkeit (englisch: accuracy) zu Berechnen:

$$accuracy = \frac{\text{Anzahl richtig klassifizierter Datensätze}}{\text{Gesamtanzahl klassifizierter Datensätze}}$$

Jedoch gibt es für den Fall, dass der Datensatz nicht richtig klassifiziert wurde, noch weitere Abstufungen, so ist der Fehler der Klassifizierung, wenn um ein Intervall daneben klassifiziert wurde (z.B. Intervall 3 statt Intervall 4) geringer, als wenn um mehrere Intervalle daneben klassifiziert wurde (z.B. Intervall 1 statt Intervall 9). Es bietet sich daher an, den Fehler F entsprechend des Abstandes der Fehlklassifizierung zu gewichten:

$$F = \frac{\sum_{i=1}^{n} |j - Bayes(a_i)|}{n}$$

mit

n Anzahl der klassifizierten Datensätze,

j Das tatsächliche Intervall des Datensatzes,

$Bayes(a_i)$ Das vom Bayes-Klassifikator vorhergesagte Intervall für den Datensatz.

Damit werden auf den ersten Blick zwar auch Datensätze betrachtet, die richtig klassifiziert wurden, diese gehen aber mit 0 in die Summe ein. Für alle weiteren Fälle gilt also; Je weiter das klassifizierte Intervall neben dem tatsächlichen Intervall liegt, desto stärker gewichtet fließt der Fehler ein.
Dieses Fehlermaß arbeitet allerdings noch mit einer Klassifikation auf das Mehrheitsintervall des vom naiven Bayes Klassifikator errechneten Zugehörigkeitshistogrammes. Es muss also ein Fehlermaß gefunden werden, welches die errechnete Verteilung berücksichtigt. Daher muss für jeden Patienten und jedes Intervall des Prognose–Histogrammes der Abstand zum tatsächlichen Intervall aufsummiert werden. Diese Summierung erfolgt gewichtet nach der errechneten Wahrscheinlichkeit des naiven Bayes Klassifikators:

$$F = \frac{\sum_{a=1}^{n} \sum_{i=1}^{k} (|j - Bayes(a_i)| \cdot p(Bayes(a_i)))}{n}$$

mit

n Anzahl aller Patienten,

k Anzahl aller Intervalle,

$Bayes(a_i)$ Vorhersageintervall für den Patienten,

j tatsächliches Intervall des Patienten,

$p(Bayes(a_i)))$ Wahrscheinlichkeit der Klassifizierung in a_k aus dem Histogramm der naiven Bayes Klassifikation.

Diese berechnete Summe wird dann durch die Anzahl der betrachteten Datensätze geteilt, um so den mittleren Fehler zu erhalten. Dieses Fehlermaß berücksichtigt die komplette errechnete Verteilung. Um die Fehlermaße für unterschiedliche Anzahlen an Diskretisierungsintervallen noch miteinander vergleichen zu können wird das Fehlermaß anschließend auf $[0, 1]$ normiert, dazu wird durch den Wert des maximalen Fehlers geteilt: $F_{norm} = \frac{F}{k-1}$.

Um die Güte des Ergebnisses bewerten zu können, bietet sich intuitiv die einheitliche Klassifizierung aller Attribute auf das Mehrheitsintervall, also jenes Intervall in welchem die meisten Datensätze liegen, an. Dieser Ansatz liefert allerdings, sofern das Mehrheitsintervall das mittlere Intervall ist, zu gute Ergebnisse. Dies liegt daran, dass der maximale Fehler bei der naiven Klassifizierung auf das mittlere Intervall nur noch halb so groß ist, wie der maximale Fehler bei der errechneten Klassifikation. Da größere Fehler auch stärker gewichtet in das Fehlermaß einfließen, hat dies zur Folge, dass mit diesem Ansatz bei einer niedrigen Anzahl an Intervallen gute Vergleichswerte errechnet werden. Je höher die Anzahl an Intervallen jedoch ist, desto unrealistischer wird dieser Wert, bis er schließlich niedriger ist, als das Fehlermaß der errechneten Klassifizierung.
Der nächste mögliche Ansatz zur Erzeugung eines Vergleichsergebnisses ist das zufällige Zuordnen eines Intervalls zu jedem Datensatz. Dieser Ansatz funktioniert auch für homogen frequent verteilte Zielintervalle uneingeschränkt gut. Die Zufälligkeit dieses Ansatzes hat allerdings zur Folge, dass eine Klassifizierung zufälligerweise viel besser oder viel schlechter als das errechnete Ergebnis sein könnte (mit steigender Größe der Datenmenge sinkt die Eintrittswahrscheinlichkeit dieser Möglichkeit). Zudem ist durch die Zufallskomponente eine deterministische Berechnung der Referenzklassifizierung nicht möglich.
Durch eine ausreichend große Anzahl an Wiederholungen der Zufallsklassifizierung würde sich das Fehlermaß letztlich dem Erwartungswert annähern. Es besteht direkt die Möglichkeit die Fehlklassifizierungswahrscheinlichkeit F_j der Datensätze eines Intervalles anzugeben. Dafür wird, anhand der Verteilungen der Datensätze, die Wahrscheinlichkeit der Fehlklassifizierung für ein Intervall berechnet. Diese ist gegeben durch:

$$F_j = \sum_{k=1}^{n} (|j-k| \cdot p_k)$$

mit

k „klassifiziertes“ Intervall,

j tatsächliches Intervall,

p_k Eintrittswahrscheinlichkeit der Datensätze im klassifizierten Intervall k zu sein.

Die Eintrittswahrscheinlichkeit wird über den relativen Anteil aller Datensätze aus der Stichprobe, die im entsprechenden Intervall liegen, geschätzt.

Dieser Fehler kann, gewichtet über den relativen Anteil der Datensätze der einzelnen Intervalle, aufsummiert werden. Unter Berücksichtigung der Eintrittswahrscheinlichkeiten erhält man dann den Erwartungswert E des Fehlermaßes:

$$E = \sum_{j=1}^{n} (\sum_{k=1}^{n} (|j-k| \cdot p_k)) \cdot p_j$$

mit

p_j Eintrittswahrscheinlichkeit der Datensätze im Intervall ($p_j = p_k$ wenn $j = k$)

Somit ist es nicht mehr notwendig, in ausreichend hoher Wiederholungsanzahl für alle Datensätze zufällig eine Klassifizierung vorzunehmen. Der Erwartungswert kann zudem deterministisch berechnet werden und gleicht dem oben beschriebenen Fehlermaß. Um die Vergleichbarkeit zu garantieren, muss wie beim Fehlermaß auf $[0, 1]$ normiert werden, indem durch den maximal vorkommenden Fehler geteilt wird: $E_{norm} = \frac{E}{k-1}$

3.5.2. K–Means Clustering

Das Fehlermaß des K–Means Clusterings ist schon im Data Mining Prozess selbst eingebaut und wird zur Berechnung der Cluster benötigt. Es handelt sich dabei um die quadrierte Summe der Fehler (SSE, siehe Abschnitt 2.1.4).

Im konkreten Fall bietet schon die Betrachtung der Kaplan–Meier–Kurven eine Abschätzung der Güte dieses Data Mining–Verfahrens zu Überlebenszeitprognose. Unterscheiden sich die Kurven deutlich voneinander, wurden durch das Clustering sinnvolle Cluster gefunden, die sich in Ihren Überlebenswahrscheinlichkeiten ebenfalls deutlich voneinander abheben.

4. Ergebnisse

4.1. Naiver Bayes Klassifikator

Die verwendeten naiven Bayes–Verfahren waren allesamt ähnlich aufgebaut. Zu Beginn steht der Import des vorverarbeiteten Exportes der Daten aus GTDS. Um den naiven Bayes Klassifikator anwenden zu können, wurde anschließend die tatsächliche Überlebenszeit diskretisiert. Dies geschah einmal äquifrequent und einmal äquidistant. Bei der Diskretisierung wurden aufgrund der nicht intuitiven Erkennbarkeit einer sinnvollen Anzahl an Intervallen mehrere Intervallanzahlen (2,3,4,5 und 10) gegeneinander getestet. Anschließend fand eine Merkmalsauswahl statt. Auch hier wurden mehrere Wege gewählt. So wurden die Backward Elimination, Forward Selection, Information Gain und die Expertenselektion gegeneinander und gegen den naiven Ansatz der No Selection getestet. Danach folgte eine zufällige Aufteilung dieses Datenbestandes in einen Trainings- und eine Testdatenmenge. Abschließend wurde mit Hilfe des naiven Bayes die Datensätze klassifiziert und das Fehlermaß F und der Erwartungswert E berechnet (siehe dazu Abbildung 4).

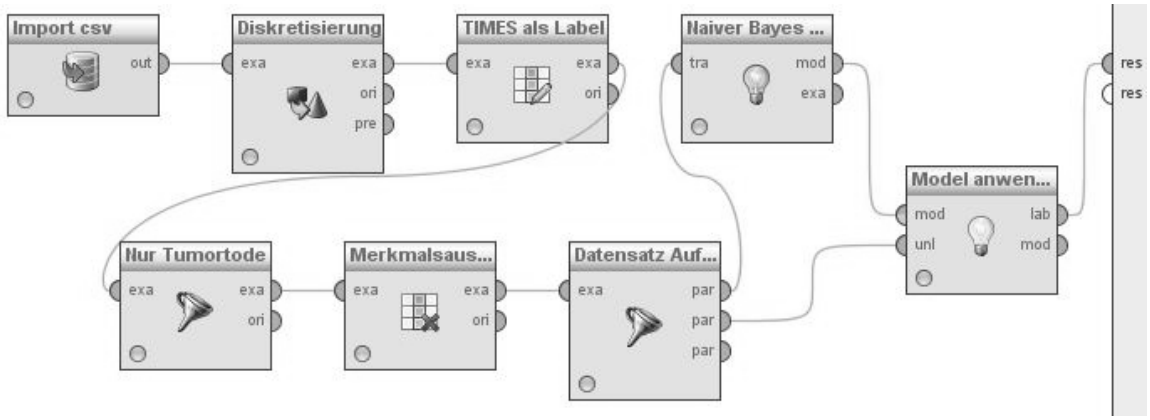

Abbildung 4: Kontrollfluss Naiver Bayes Klassifikator

Wie in der Abbildung 4 zu sehen ist, wurden zusätzlich zwei weitere Schritte vorgenommen. Zum einen wurde die Überlebenszeit als das Attribut auf dem die Klassifizierung auszuführen ist gesetzt, zum anderen wurden alle Datensätze die keinen Tumortod enthielten gefiltert, da der naive Bayes Klassifikator keine Möglichkeit bietet die Zensuren adäquat zu behandeln. Die genauen Prozesse weichen teilweise von dem oben Dargestellten ab, alle Prozesse sind daher im Begleitmaterial zu der Arbeit zu finden.

Nachfolgend sind die Ergebnistabellen der normierten Fehler und Erwartungswerte nach Merkmalsauswahlverfahren sortiert aufgelistet. Die Auflistung welches Merkmalsauswahlverfahren, welche Attribute ausgewählt hat ist aus Gründen der Übersichtlichkeit im Anhang A in den Tabellen 20 (für die äquidistante Diskretisierung) und 21 (für die äquifrequente Diskretisierung) zu finden.

Aus Platzgründen werden im folgenden Abkürzungen verwendet, hierzu die Legende:

NS: No Selection

BE: Backward Elimination

FS: Forward Selection

IG: Information Gain

ES: Expert Selection

EF: Äquifrequente Diskretisierung

ED: Äquidistante Diskretisierung

F(X): Fehlermaß von X

E(X): Erwartungswert für X

Da für jede Versuchsdurchführung ein Fehlermaß und ein Erwartungswert berechnet wurde, wurde am Ende ein Mittelwert über alle Fehlermaße bzw. Erwartungswerte eines Verfahrens berechnet, um eine einfache Vergleichs–Möglichkeit zu erhalten.

4.1.1. Naiver Ansatz

Die Fehlermaße und Erwartungswerte des naiven Bayes Klassifikator ohne vorherige Merkmalsauswahl sind in Tabelle 4 dargestellt.

n Intervalle	F(NS_EF)	E(NS_EF)	F(NS_ED)	E(NS_ED)
2	0.39	0.49	0.15	0.03
3	0.34	0.41	0.17	0.04
4	0.31	0.38	0.22	0.05
5	0.31	0.37	0.2	0.06
10	0.32	0.34	0.17	0.08
Mittelwert	0.334	0.398	0.182	0.052

Tabelle 4: No Selection Ergebnistabelle

4.1.2. Backward Elimination

Es wurden durch die Backward Elimination zwischen 4 und 12 Attribute ausgewählt. Die naive Bayes Klassifikation mit vorhergehender Backward Elimination lieferte die in Tabelle 5 aufgelisteten Ergebnisse:

n Intervalle	F(BE_EF)	E(BE_EF)	F(BE_ED)	E(BE_ED)
2	0.39	0.49	0.02	0.03
3	0.33	0.42	0.04	0.03
4	0.3	0.38	0.05	0.05
5	0.31	0.36	0.05	0.06
10	0.26	0.33	0.08	0.09
Mittelwert	0.318	0.396	0.048	0.052

Tabelle 5: Backward Elimination Ergebnistabelle

4.1.3. Forward Selection

Ebenso wie die Backward Elimination konnte bei der Forward Selection von Anwender in Rapid Miner keine konkrete Anzahl an Attributen vorgegeben werden, die vom Selektionsalgorithmus erreicht werden sollte, es wurden von der Forward Selection zwischen 1 und 7 Attributen ausgewählt. Diese beiden Verfahren (Backward Elimination und Forward Selection) am besten miteinander zu vergleichen, da vom Anwender keine Einflussmöglichkeit auf die Anzahl der Attribute bestand. Mit der Forward Selection erzielte der naive Bayes Klassifikator folgende, in Tabelle 6 aufgelisteten, Ergebnisse:

n Intervalle	F(FS_EF)	E(FS_EF)	F(FS_ED)	E(FS_ED)
2	0.4	0.49	0.03	0.04
3	0.36	0.42	0.04	0.04
4	0.33	0.39	0.05	0.05
5	0.31	0.38	0.07	0.07
10	0.31	0.35	0.09	0.09
Mittelwert	0.342	0.406	0.056	0.058

Tabelle 6: Forward Selection Ergebnistabelle

4.1.4. Expertenauswahl der Attribute

Die Expertenselektion wurde von Fr. Dr. Bochum durchgeführt. Dabei wurden die Attribute in Klassen gleicher Aussagekraft eingeteilt. Die Expertenselektion wurde mehrmals durchgeführt, wobei jedes mal ein weitere dieser Klasse dazugenommen wurde. Die Ergebnisse dieser Versuchsdurchführungen sind in den Tabellen 7 (für 2 Attribute), 8 (für 7 Attribute), 9 (für 8 Attribute) und 10 (für 10 Attribute) dargestellt.

n Intervalle	F(ES2_EF)	E(ES2_EF)	F(ES2_ED)	E(ES2_ED)
2	0.48	0.49	0.14	0
3	0.4	0.41	0.13	0.01
4	0.36	0.38	0.11	0.02
5	0.35	0.37	0.18	0.04
10	0.33	0.34	0.19	0.07
Mittelwert	0.384	0.398	0.15	0.028

Tabelle 7: Expertenauswahl Ergebnistabelle für 2 Attribute

n Intervalle	F(ES7_EF)	E(ES7_EF)	F(ES7_ED)	E(ES7_ED)
2	0.43	0.49	0.33	0
3	0.38	0.41	0.4	0.01
4	0.35	0.38	0.33	0.02
5	0.35	0.37	0.27	0.04
10	0.35	0.34	0.25	0.07
Mittelwert	0.372	0.398	0.316	0.028

Tabelle 8: Expertenauswahl Ergebnistabelle für 7 Attribute

n Intervalle	F(ES8_EF)	E(ES8_EF)	F(ES8_ED)	E(ES8_ED)
2	0.43	0.49	0.33	0
3	0.38	0.41	0.36	0.01
4	0.34	0.38	0.032	0.02
5	0.34	0.37	0.26	0.04
10	0.34	0.34	0.24	0.07
Mittelwert	0.366	0.398	0.2444	0.028

Tabelle 9: Expertenauswahl Ergebnistabelle für 8 Attribute

n Intervalle	F(ES10_EF)	E(ES10_EF)	F(ES10_ED)	E(ES10_ED)
2	0.41	0.49	0.21	0
3	0.37	0.41	0.31	0.01
4	0.33	0.38	0.25	0.02
5	0.33	0.37	0.21	0.04
10	0.34	0.34	0.19	0.07
Mittelwert	0.356	0.398	0.234	0.028

Tabelle 10: Expertenauswahl Ergebnistabelle für 10 Attribute

4.1.5. Information Gain

Um eine Vergleichbarkeit zur Expertenselektion herzustellen, wurde die Gewichtung des Information Gain–Verfahrens jeweils auf die gleichen Mengen beschränkt, wie es die Selektion von

Fr. Bochum vorgab. Die Ergebnisse dazu sind in den Tabellen 11 (für 2 Attribute), 12 (für 7 Attribute), 13 (für 8 Attribute) und 14 (für 10 Attribute) aufgelistet.

n Intervalle	F(IG2_EF)	E(IG2_EF)	F(IG2_ED)	E(IG2_ED)
2	0.45	0.49	0	0
3	0.37	0.4	0.02	0.01
4	0.33	0.37	0.08	0.02
5	0.32	0.36	0.12	0.04
10	0.29	0.32	0.13	0.07
Mittelwert	0.352	0.388	0.07	0.028

Tabelle 11: Information Gain Ergebnistabelle für 2 Attribute

n Intervalle	F(IG7_EF)	E(IG7_EF)	F(IG7_ED)	E(IG7_ED)
2	0.41	0.49	0	0
3	0.34	0.4	0.08	0.01
4	0.31	0.37	0.06	0.02
5	0.3	0.36	0.07	0.04
10	0.28	0.32	0.09	0.07
Mittelwert	0.328	0.388	0.06	0.028

Tabelle 12: Information Gain Ergebnistabelle für 7 Attribute

n Intervalle	F(IG8_EF)	E(IG8_EF)	F(IG8_ED)	E(IG8_ED)
2	0.41	0.49	0	0
3	0.34	0.4	0.02	0.01
4	0.31	0.37	0.04	0.02
5	0.3	0.36	0.09	0.04
10	0.28	0.32	0.09	0.07
Mittelwert	0.328	0.388	0.048	0.028

Tabelle 13: Information Gain Ergebnistabelle für 8 Attribute

n Intervalle	F(IG10_EF)	E(IG10_EF)	F(IG10_ED)	E(IG10_ED)
2	0.4	0.49	0	0
3	0.34	0.4	0.13	0.01
4	0.31	0.37	0.04	0.02
5	0.3	0.36	0.16	0.04
10	0.27	0.32	0.07	0.07
Mittelwert	0.324	0.388	0.08	0.028

Tabelle 14: Information Gain Ergebnistabelle für 10 Attribute

4.2. K–Means Clustering

Für das K–Means Clustering wurde zuerst ein k zu mittlerer Abstand Diagramm für $k = 2..30$ erstellt, der Kontrollfluss dazu ist in Abbildung 5 dargestellt:

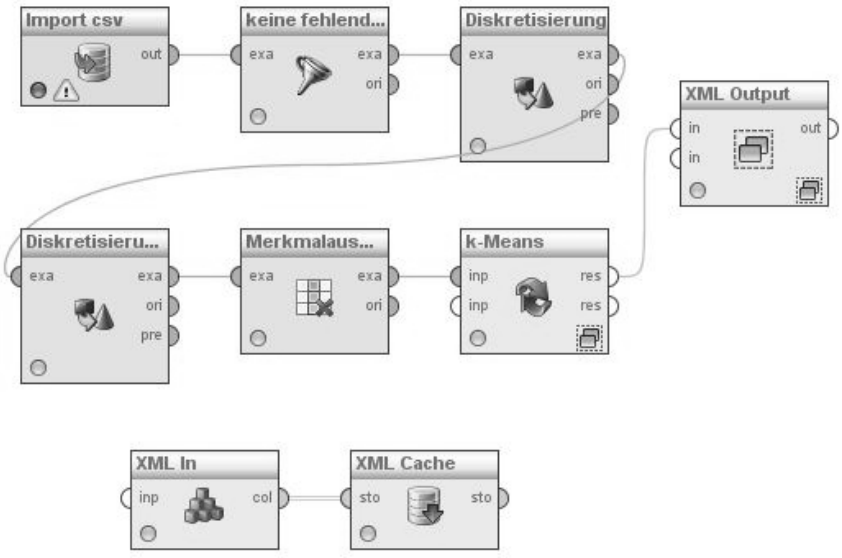

Abbildung 5: Kontrollfluss Clustering Auswahl k

Dabei wird zuerst die Comma Separated Value (csv)-Datei importiert. Anschließen werden Datensätze mit fehlenden Werten herausgefiltert, da das Clustering nicht mit fehlenden Werten umgehen kann. Zu beachten ist dabei, dass eine Vorverarbeitung der Daten zu diesem Zeitpunkt schon stattgefunden hat, also nur noch wenige Werte fehlen. Konkret handelt es sich dabei um Datensätze für die kein UICC-Staging bestimmt werden konnte, da kein Staging dokumentiert war. Anschließend werden einige Attribute (allerdings nicht die Überlebenszeit) diskretisiert. Danach folgt die Merkmalsauswahl, bei welcher auch die Überlebenszeit herausgefiltert wird. Abschließend wird die Menge der Datensätze geclustered. Dies geschieht in einer Schleife für alle k Werte aus $[2, 30]$. Zusammen mit der mittleren Distanz der einzelnen Punkte zu ihrem Clusterzentrum wird das Ergebnis in eine XML-Datei geschrieben, welche später ausgelesen wird um aus ihr das benötigte Diagramm zu generieren. Aus diesem Diagramm wurden die Kanten abgelesen (ggf. musste die bessere Kante zuvor errechnet werden) und mit dieser Zahl an Clusterzentren das eigentliche Clustering durchgeführt, dieser Prozess gleicht dem in Abbildung 5, allerdings wird, wie in Abbildung 6 zu sehen ist, das Clustering nur noch einmalig für den zuvor herausgefundenen k-Wert durchgeführt:

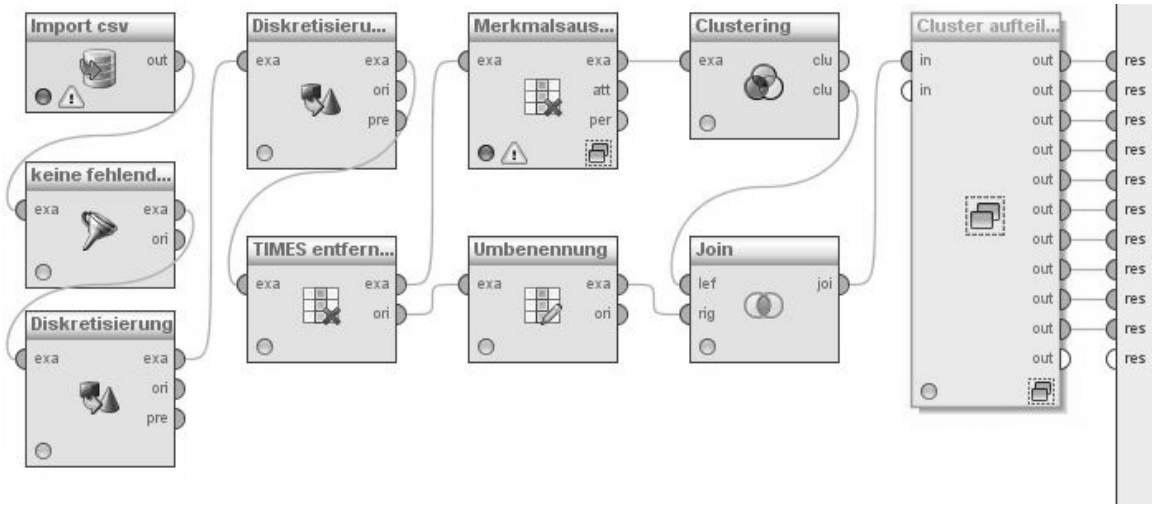

Abbildung 6: Kontrollfluss Clustering

Hierbei ist zu beachten, dass die Überlebenszeit für das Clustering aus der zu clusternden Menge der Datensätze entfernt und später wieder hinzugefügt wird. Abschließend wird der komplette Datenbestand in den einzelnen errechneten Clustern ausgegeben. Aus dieser Ausgabe müssen dann nur noch die Clusterzentren abgelesen und Kaplan–Meier–Schätzer generiert werden.

Welche Attribute von den Merkmalsauswahlverfahren jeweils ausgewählt wurden ist der Tabelle 19 in Anhang A zu entnehmen. Hier folgt die Darstellung der Kaplan–Meier–Schätzer sortiert nach den Merkmalsauswahlverfahren. Die k-Diagramme sind jeweils direkt davor zu finden. Die Kaplan–Meier–Kurven werden unschärfer je weiter die Zeit fortschreitet, da immer weniger Patienten betrachtet werden. Aussagekräftig ist daher vor allem der vordere Teil des Kaplan–Meier–Diagrammes (ca. bis Tag 2000).

Der Übersichtlichkeit halber wurde auch hier für die Darstellung der Kaplan–Meier–Diagramme der Datensatz mit der längsten Überlebenszeit nicht mit einbezogen. Dadurch wurde ein Stauchen aller Kurven vermieden.

4.2.1. Naiver Ansatz

Da k-Diagramm für den naiven Ansatz stellte sich wie in Abbildung 7 dar.

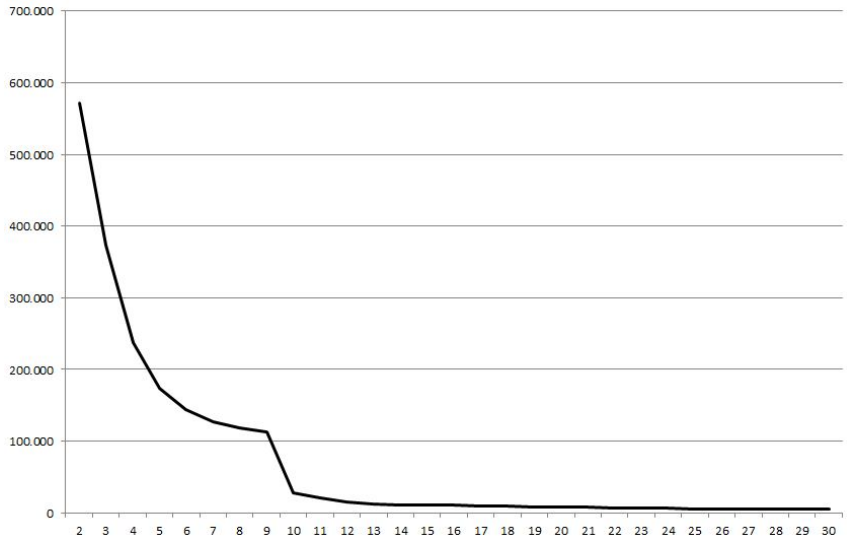

Abbildung 7: k-Diagramm Naiver Ansatz

Dabei ist eine deutliche Kante für $k = 10$ Attribute zu erkennen. Ein Cluster enthielt dabei allerdings nur 13 Datensätze (Cluster_4), weshalb es im Kaplan–Meier–Diagramm nicht gezeichnet wurde. Alle anderen Kaplan–Meier–Kurven sind hier (Abbildung 8) dargestellt:

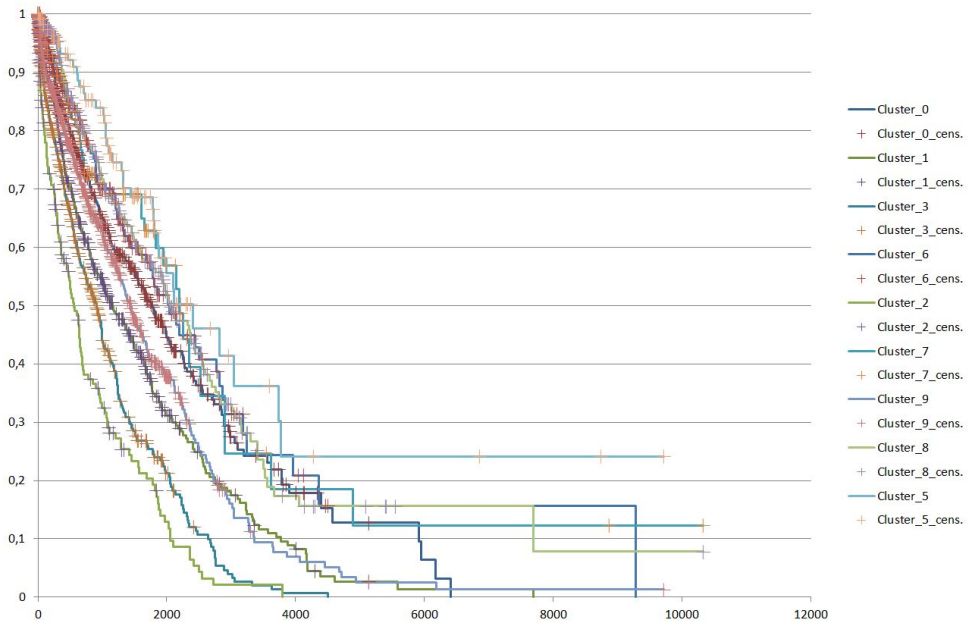

Abbildung 8: Kaplan Meier-Kurve für No Selection

Trotz dieses naiven Ansatzes können die Kaplan–Meier–Schätzer gut voneinander abgegrenzt werden.

4.2.2. Backward Elimination

Für das K–Means Clustering mit Backward Elimination als Merkmalsauswahlverfahren ergab sich das in Abbildung 9 dargestellte k-Diagramm.

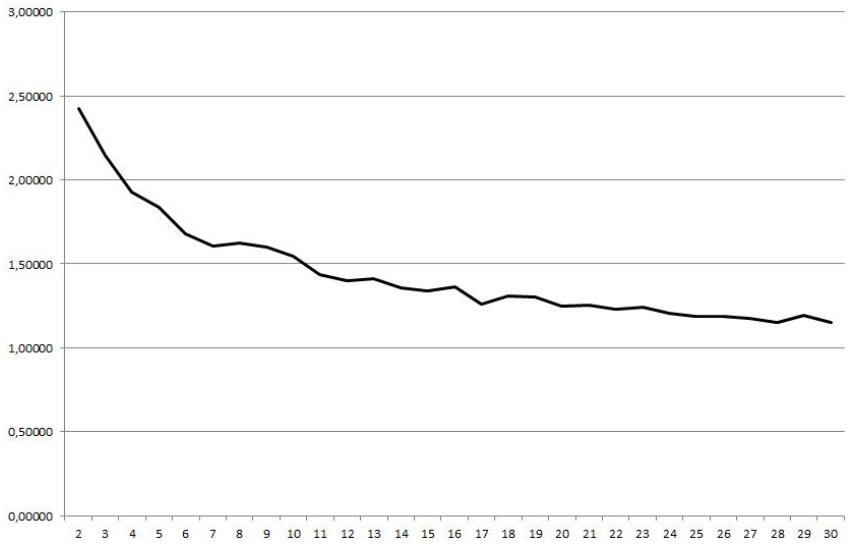

Abbildung 9: k-Diagramm Backward Elimination

Die stärkste Kante ist deutlich bei $k = 7$ Attributen zu erkennen. Bei $k = 7$ wurden von der Backward Elimination 15 relevante Attribute gefunden. Dafür ergibt sich dann des Kaplan–Meier–Diagramm aus Abbildung 10

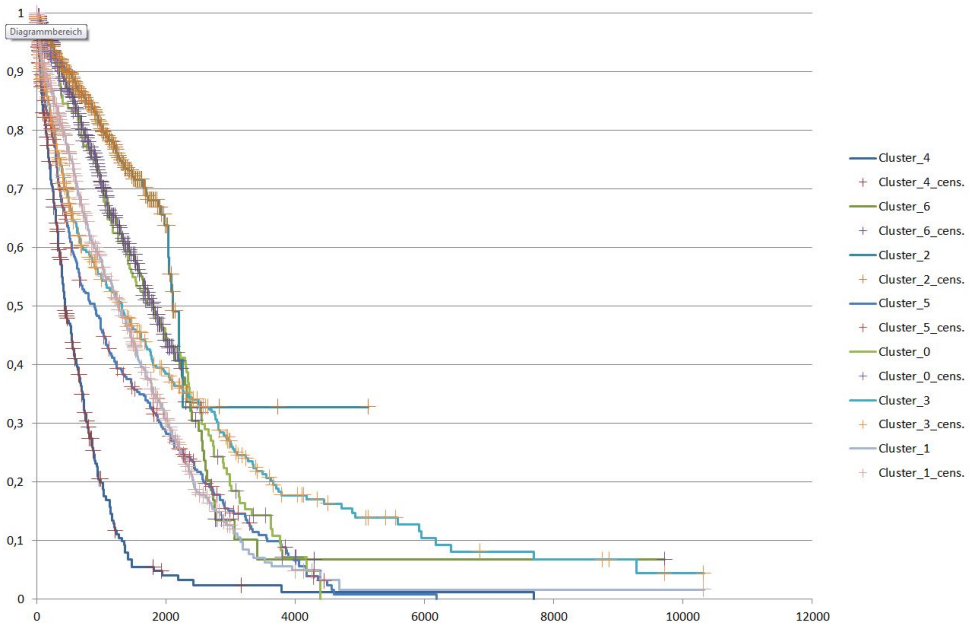

Abbildung 10: Kaplan Meier-Kurve für Backward Elimination

4.2.3. Forward Selection

Das k-Diagramm der Forward Selection stellte sich wie in Abbildung 11 dar.

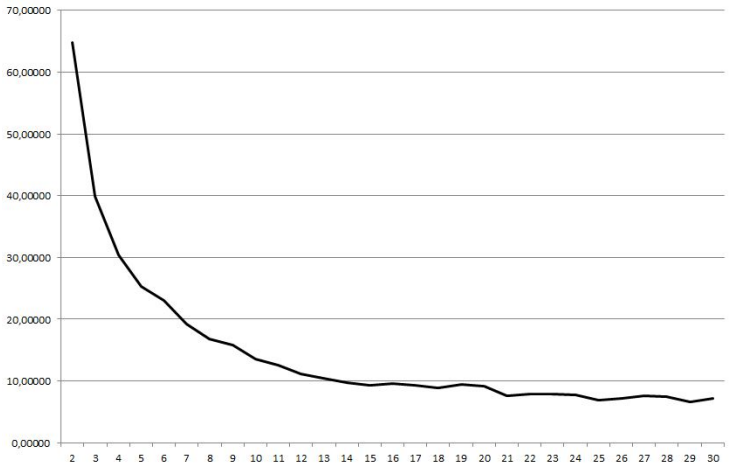

Abbildung 11: k-Diagramm Forward Selection

Dabei sind zwei 3 potentielle Kanten an den Stellen 5, 8 und 10 auszumachen, alle weiteren potentiellen Kanten sind auf zufällige Schwankungen zurückzuführen. Es ergaben sich für diese Stellen folgende Δx Werte: $1,44$ (bei $x = 5$), $0,74$ (bei $x = 8$) und $0,59$ (bei $x = 10$). Daher wurde die Stelle $x = 5$ für das Clustering verwendet. Es wurden an dieser Stelle vom Forward Selecion Algorithmus 1 relevante Attribute ausgemacht. Er ergibt sich das in Abbildung 12 dargestellt Kaplan–Meier–Diagramm.

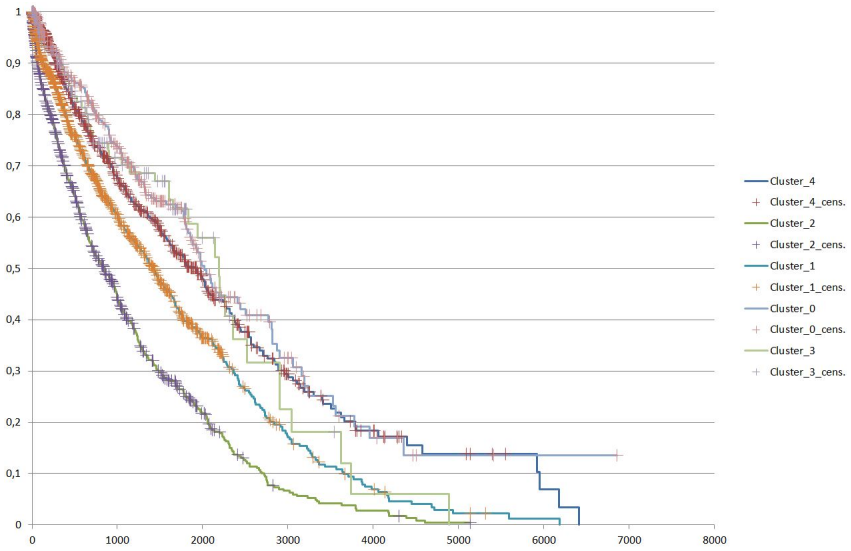

Abbildung 12: Kaplan Meier-Kurve für Forward Selection

Die Kaplan–Meier–Kurven sind hier ebenfalls fast alle klar voneinander abzugrenzen.

4.2.4. Expertenauswahl der Attribute

Die Expertenauswahl erfolgte für 2,7,8 und 10 Attribute.

Das k-Diagramm für die Expertenselektion (siehe Abbildung 13) mit auf 2 Attribute stellt sich nicht so gleichförmig da, wie alle vorherigen k-Diagramme.

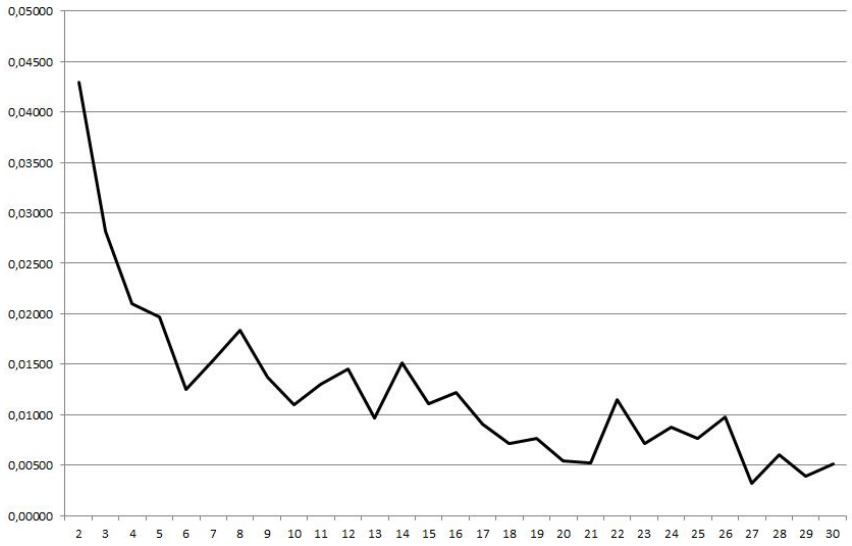

Abbildung 13: k-Diagramm Expertenselektion auf 2 Attribute

Es findet sich die stärkste Kante bei $k = 6$, dafür ergibt sich das Kaplan–Meier–Diagramm aus Abbildung 14.

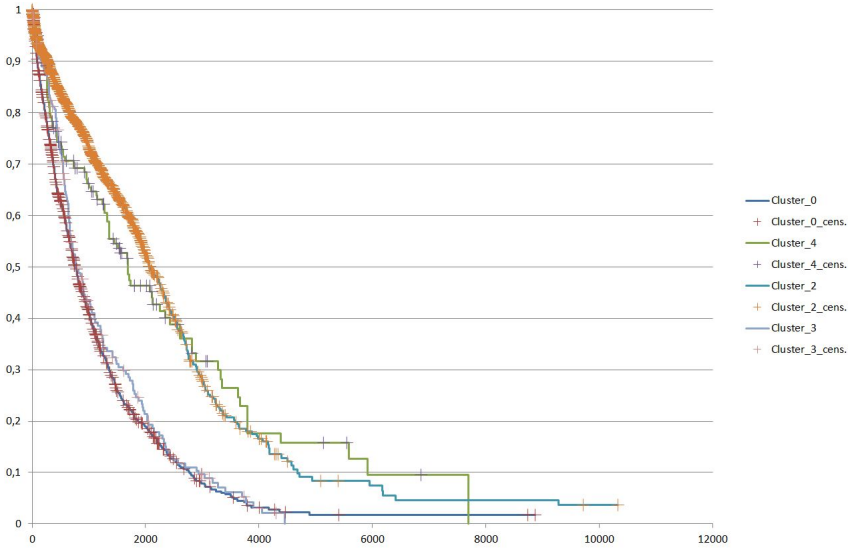

Abbildung 14: Kaplan–Meier–Diagramm für Expertenselektion mit 2 Attributen

Ein Cluster wurde dabei aufgrund zu weniger Attribute nicht gezeichnet. Es ist zudem zu erkennen, dass jeweils zwei von den verbleibenden vier Clustern sehr ähnliche (fast gleiche) Verläufe der Kaplan–Meier–Schätzer aufzuweisen haben.

Das k-Diagramm für 7 Attribute (siehe Abbildung 15) hat seine stärkste Kante bei $k = 7$.

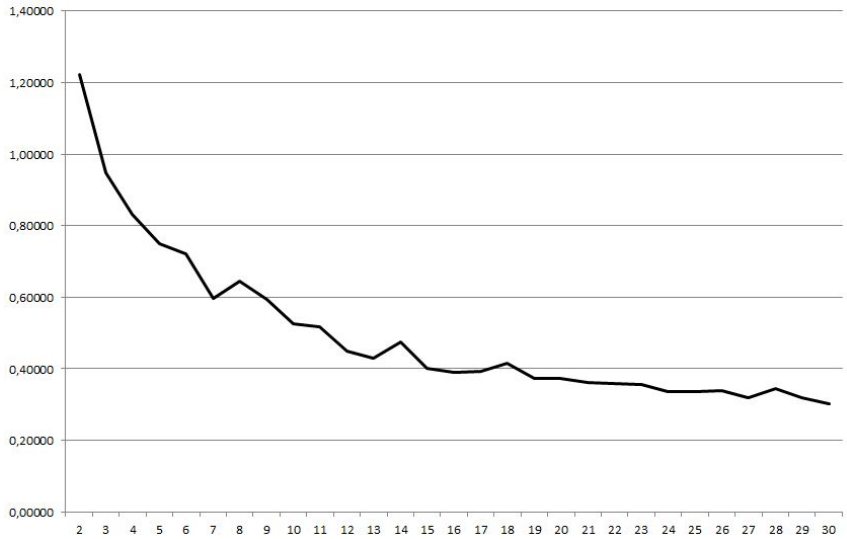

Abbildung 15: k-Diagramm Expertenselektion auf 7 Attribute

Für sieben Cluster ergibt sich folgendes Kaplan–Meier–Diagramm aus Abbildung 16.

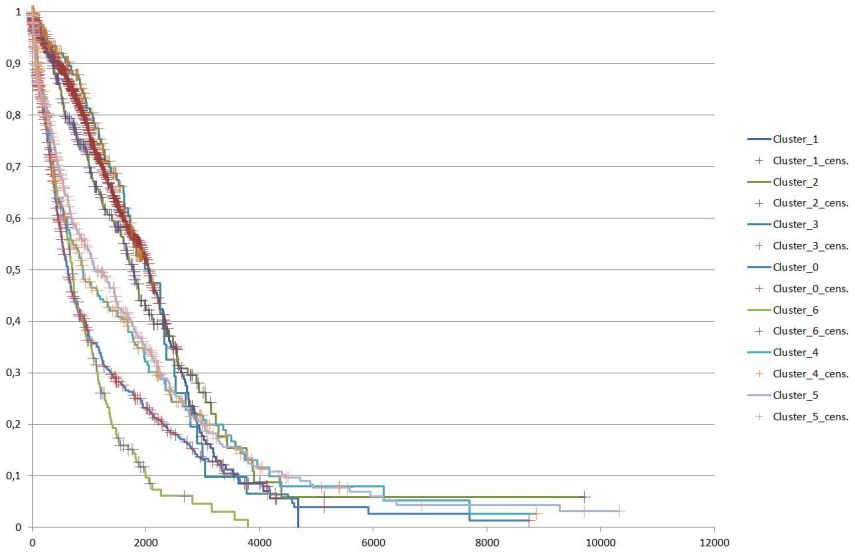

Abbildung 16: Kaplan–Meier–Diagramm für Expertenselektion mit 7 Attributen

Die Kaplan–Meier–Schätzer sind hier je Cluster nur schwer voneinander zu trennen. die 7 Cluster teilen sich wieder in zwei Gruppen auf, wovon eine leicht längere Überlebenszeiten haben.

Im k-Diagramm für 8 Attribute (siehe Abbildung 17) konnten zwar mehrere Kanten visuell ausgemacht werden. Alle gefundenen Kanten wiesen jedoch so hohe Anzahlen an Clustern auf, dass keine vernünftigen Kaplan–Meier–Diagramme mehr gezeichnet werden konnten.

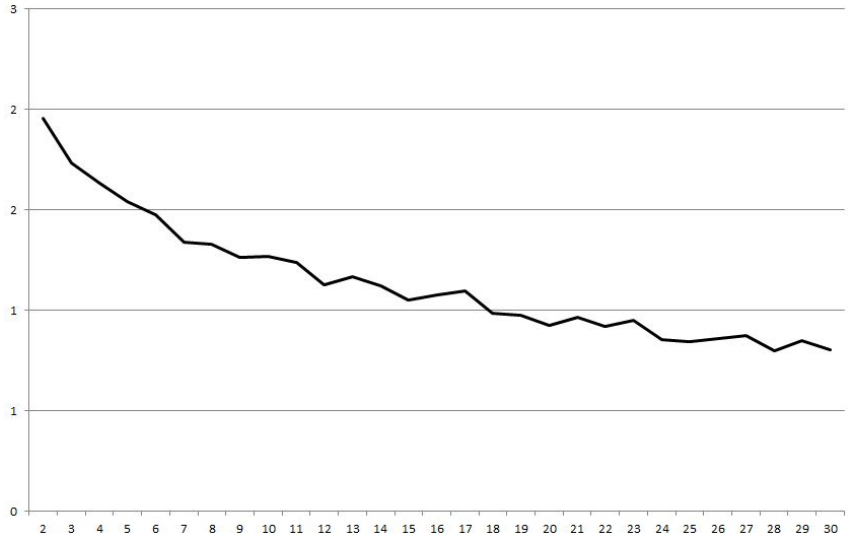

Abbildung 17: k-Diagramm Expertenselektion auf 8 Attribute

Das k-Diagramm für 10 Attribute ist in Abbildung 18 dargestellt.

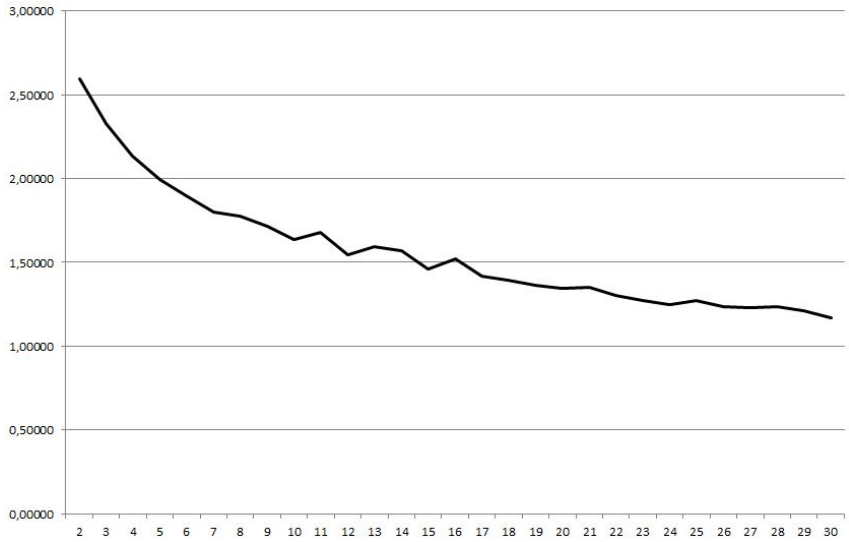

Abbildung 18: k-Diagramm Expertenselektion auf 10 Attribute

Für 10 Attribute fanden sich 3 potentielle Kanten bei $x = 10, 12, 15$. An diesen Stellen wurden folgende folgende Δx Werte berechnet: $0,06$ (bei $x = 10$), $0,04$ (bei $x = 12$) und $0,06$ (bei $x = 15$). Da die Kanten bei $x = 10$ und $x = 15$ scheinbar ähnlich stark zu sein scheinen, wurde entschieden das Kaplan–Meier–Diagramm für die niedrigere Anzahl (also 10 Cluster) zu

generieren — siehe Abbildung 19

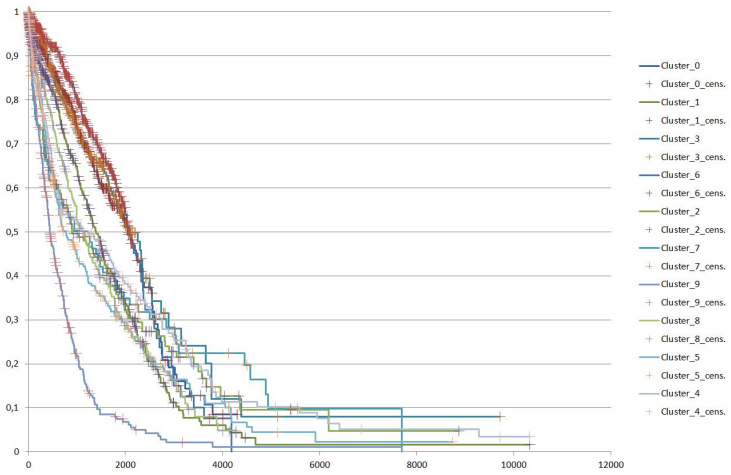

Abbildung 19: Kaplan–Meier–Diagramm für Expertenselektion mit 10 Attributen

Aufgrund der hohen Anzahl an Kurven in einem Diagramm, sind einzelne Kurven nur noch schwer voneinander zu unterscheiden. Es fällt jedoch auf, dass sich nur eine Kurve (von Cluster 4) wesentlich von allen anderen abhebt. Die anderen Kurven teilen sich jedoch nochmals, allerdings nicht so deutlich, in zwei Gruppen mit leicht unterschiedlichen Überlebenszeiten.

4.2.5. Information Gain

Entsprechend der Expertenselektion fand wurde das Clustering Verfahren mit den 2,7,8 und 10 besten Attributen des Information Gain Ansatzes durchgeführt. Das k-Diagramm für 2 Attribute ist in Abbildung 20 dargestellt.

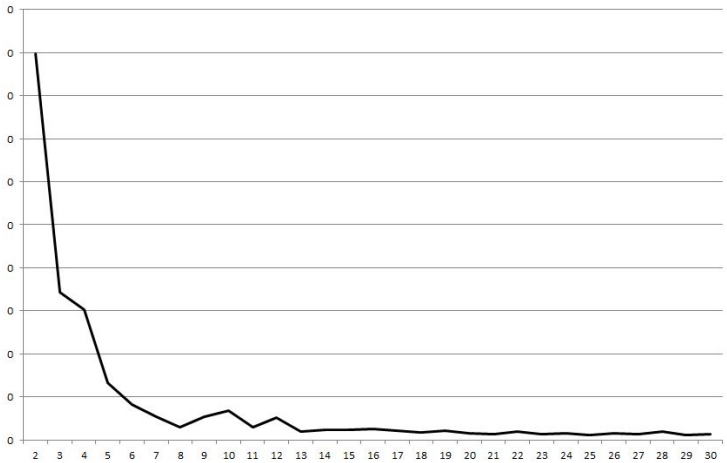

Abbildung 20: k-Diagramm Infomation Gain mit 2 Attribute

Für die deutlich zu erkennende stärkste Kante bei $x = 8$ ist das Kaplan–Meier–Diagramm in Abbildung 21 gezeichnet worden.

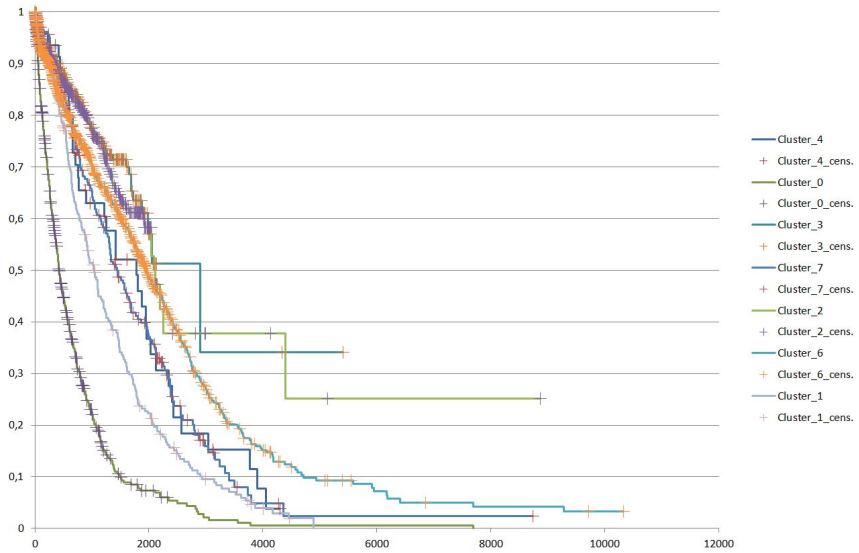

Abbildung 21: Kaplan–Meier–Diagramm für Information Gain mit 2 Attributen

Für ein Cluster wurden beim K–Means Clustering keine passenden Elemente gefunden. Zudem enthalten manche der Kurven nur sehr wenige Datensätze. Dadurch können die teilweise sehr starken Sprünge in den Kurven erklärt werden. Alle restlichen Kurven lassen sich gut voneinander abgrenzen.

Das k-Diagramm für 7 Attribute ist in Abbildung 22 dargestellt.

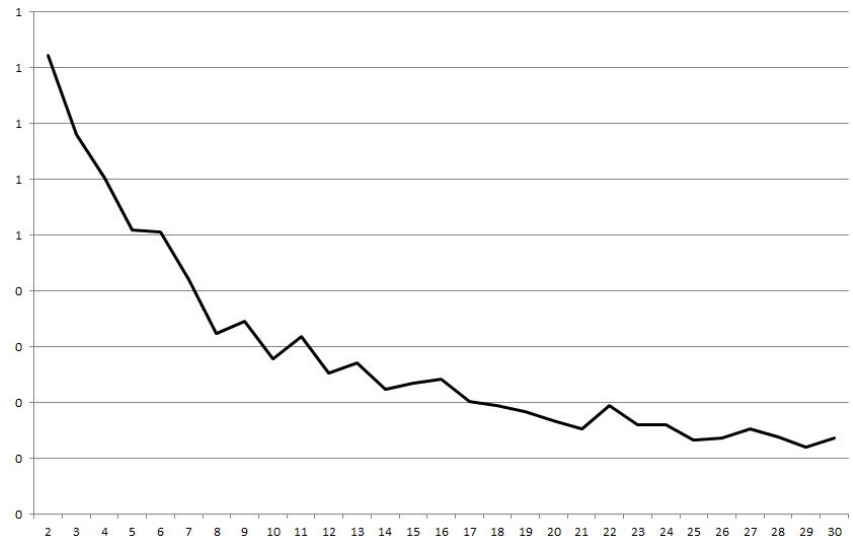

Abbildung 22: k-Diagramm Infomation Gain mit 7 Attribute

Für 7 Attribute fanden sich 2 potentielle Kanten bei $x = 8, 10$. An diesen Stellen wurden folgende folgende Δx Werte berechnet: $0, 06$ (bei $x = 8$) und $0, 055$ (bei $x = 10$). Daher wurde das Kaplan–Meier–Diagramm für 8 Cluster in Abbildung 23 gezeichnet.

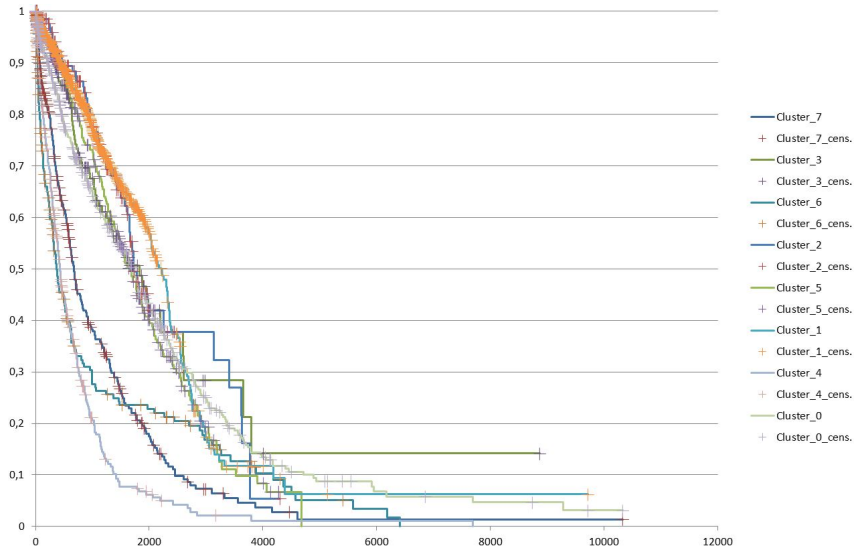

Abbildung 23: Kaplan–Meier–Diagramm für Information Gain mit 7 Attributen

Diese 8 Schätzer teilen sich im wesentlichen in 2 Gruppen, wovon eine eine deutlich bessere Überlebenszeit aufzuweisen hat.

Das k-Diagramm für 8 Attribute ist in Abbildung 24 dargestellt.

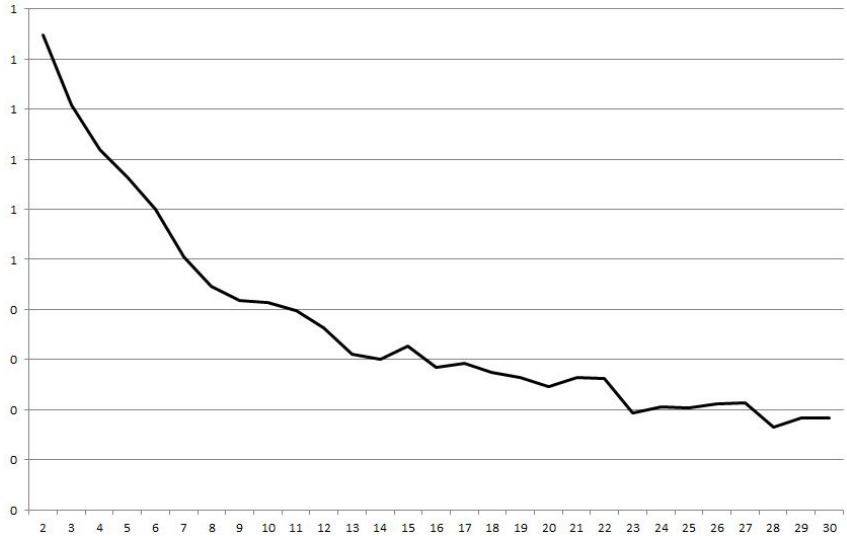

Abbildung 24: k-Diagramm Infomation Gain mit 8 Attribute

Hier konnte wieder keine Kante für ein außreichend niedriges k ausgemacht werden, sodass

sich ein sinnvolles Kaplan–Meier–Diagramm zeichnen ließe. Daher wurde kein Kaplan–Meier–Diagramm für das Clustering mit Information Gain bei 8 Attributen angefertigt.

Das k-Diagramm für 10 Attribute ist in Abbildung 25 dargestellt.

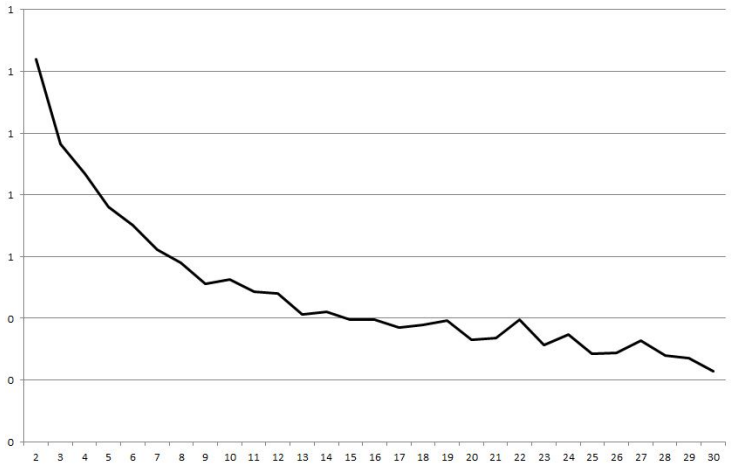

Abbildung 25: k-Diagramm Infomation Gain mit 10 Attribute

Für die deutlich zu erkennende stärkste Kante bei $x = 9$ ist das Kaplan–Meier–Diagramm in Abbildung 26 gezeichnet.

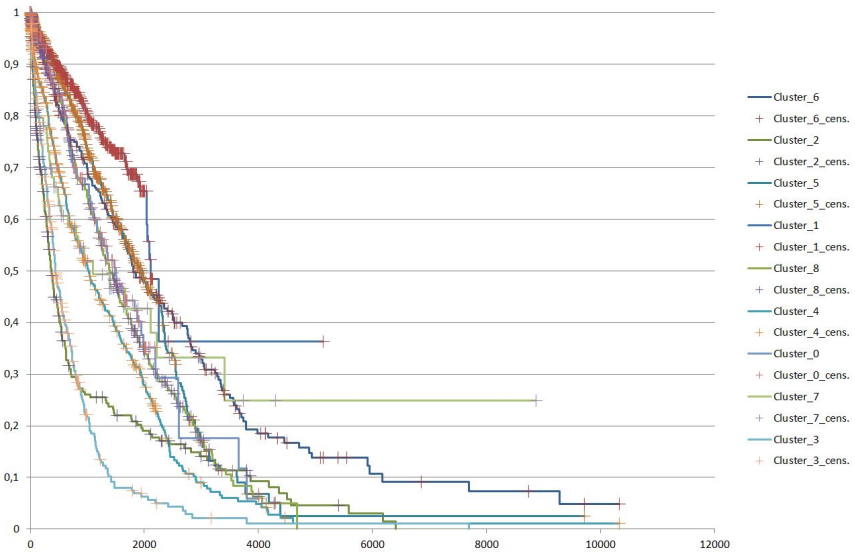

Abbildung 26: Kaplan–Meier–Diagramm für Information Gain mit 10 Attributen

In diesem Diagramm sind ebenfalls wieder viele Kurven vorhanden, was ein genaues Ablesen erschwert. Allerdings können die vorhandenen Kurven dennoch verhältnismäßig gut voneinander abgegrenzt werden.

5. Diskussion

5.1. Verfahren

Im Laufe der Erarbeitung waren trotz des praktischen Data Mining Werkzeuges viel menschliche Interaktion notwendig, so mussten beispielsweise alle 110 Durchläufe der naiven Bayes Klassifikation von Hand gestartet und die Auswertung semiautomatisch vorgenommen werden. Bei dieser Menge an menschlicher Interaktion sind Fehler nicht auszuschließen, wünschenswert wären mehr Möglichkeiten zur Automatisierung gewesen um das menschliche Fehlerpotential zu minimieren und die zur Verfügung stehende Zeit zum Verbesserung, Entwicklung und weiteren Auswertung der Data Mining–Prozessen nutzen zu können, statt die Zeit für Stapelarbeit aufwenden zu müssen.

5.2. Ergebnisse

Eine für Data Mining–Verfahren verhältnismäßig niedrige Anzahl an Datensätzen (3759 Stück) ermöglichte zwar das Testen vieler verschiedener Verfahren, aufgrund der niedrigen Laufzeiten der einzelnen Durchläufe. Auf der anderen Seite bedeutet dies aber auch, dass die Ergebnisse nicht so gut getestet und validiert werden konnten, da die Anzahl der Datensätze dafür zu gering war.

5.2.1. naiver Bayes Klassifikator

Der naive Bayes Klassifikator wurde in insgesamt 110 verschiedenen Versuchkonstellationen verwendet. Die dabei eingebauten Merkmalsauswahlverfahren berechneten jeweils unterschiedliche Attributmengen als relevant. Wie häufig ein Attribut als relevant angesehen wurde ist in Tabelle 15 dargestellt. In der Tabelle nicht berücksichtigt ist der naive Ansatz der Merkmalsauswahl, da dort immer alle Attribute ausgewält wurden.

Attributname	Vorkommen in %
ERSTE_R_KLASSIFIKATION	73
ANZAHL_METASTASEN	69
UICC	65
ANZAHL_TUMOREN	59
PRIMAERTHERAPIE	51
OP_INTENTION	47
ICD10	46
HISTO_GRADING	46
ARZT_ANLASS	41
P_L	37
ALTER	32
P_V	30
ERFASSUNGSANLASS	25
BEHANDLUNGSANLASS	23
GESCHLECHT	13

Tabelle 15: Häufigkeit der Verwendung der Attribute beim K–Means Clustering

Kein Attribut wurde in jeder Versuchskonstellation als relevant erachtet. Das Diagramm muss in so fern kritisch betrachtet werden, da durch die mehrfache Versuchsdurchfürung auf Basis der Expertenselektion (für die besten 2,7,8 und 10 Attribute) als Merkmalsverfahren die von Frau Bochum ausgewählten Attribute sehr stark gewichtet in die Häufigkeiten mit einbezogen werden. Dennoch ist ein deutlicher Unterschied in der Relevanz der einzelnen Attribute zu erkennen, was auf die höhere informationstechnische Bedeutung einzelner Attribute schließen lässt.

Für die naive Bayes Klassifikation erwies sich die Diskretisierung in äquifrequente Intervalle als fast durchweg besser. Zwar ist das Fehlermaß der äquidistanten Diskretisierung immer kleiner. Jedoch hebt sich das Fehlermaß der äquidistanten nie stark vom Erwartungswert ab, oder ist in vielen Fällen sogar schlechter als der Erwartungswert. Dies kann durch die kleine Menge an Daten erklärt werden mit denen der naive Bayes Klassifikator trainiert wurde. Da vor bei der äquidistanten Diskretisierung vor allem in den hinteren Intervallen nur wenige Attribute waren, wurde der Klassifikator nach der zufälligen Aufteilung zufälliger Weise auf diese übertrainiert. Mit einer größeren Datenmenge würde dise Abweichung sicherlich geringer ausfallen.
Dennoch muss festgehalten werden, dass bei der gegebenen Konstellation die äquidistanten Diskretisierung deutlich schlechtere Ergebnisse als die äquifrequente Diskretisierung lieferte.

Es wurden auch verschiedene Anzahlen an Diskretisierungsintervallen gegeneinander getestet. Dabei ist nach der Normalisierung auf den maximalen Fehler auf den ersten Blick ein einfacher Zusammenhang festzustellen: „Je mehr Intervalle, desto kleiner ist der Fehler“. Bei genauerer Betrachtung ist jedoch zu erkennen, dass das Fehlermaß bei einigen Verfahren für 10 Zielintervalle größer ist als das vergleichbare Fehlermaß für 5 Diskretisierungsintervalle.
Eine optimale Anzahl von Diskretisierungsintervallen kann also nicht pauschal festgelegt werden, sondern muss für jedes Merkmalsauswahlverfahren individuell gefunden werden. Diese Anzahl liegt bei den hier durchgeführten Versuchsreihen meistens bei 5 Intervallen. Die Ergebnisse der Backward Elimination suggerieren, dass die optimale Anzahl an Diskretisierungsintervallen dort

größer als 10 sein sollte.

Um zu bestimmen welche Merkmalsauswahlverfahren sich für die Fragestellung dieser Thesis eignen, muss verglichen werden welche Ergebnisse im Mittel besser sind als die des naiven Vergleichsansatzes. Diese Kriterium erfüllen nur die Backward Elimination und das Information Gain–Verfahren mit 7,8 und 10 Attributen. Es ist auch hier folgender Zusammenhang zu sehen; Verfahren, die mehr (aber nicht alle) Attribute berücksichtigen liefern bessere Ergebnisse. So ist die Forward Selection, die teilweise nur ein Attribut als relevant erachtet hat deutlich schlechter als der naive Ansatz. Auch bei der Expertenselektion und dem Information Gain–Verfahren kann erkannt werden, dass die Ergebnisse besser werden je mehr Attribute berücksichtigt werden.
Es überrascht, dass die Expertenselektion schlechter abschneidet als der naive Ansatz. Für 2 Attribute kann dies noch mit dem oben beschriebenen Zusammenhang erklärt werden. In den anderen Fällen bedeutet dieses Ergebnis allerdings, dass die ausgewählten Attribute von medizinischer Relevanz für die Behandlung sind, da sie von einer Expertin ausgewählt wurden. Allerdings gibt es Attribute, die unabhängig von der medizinischer Bedeutung einen starken Einfluss auf die Überlebenszeit haben. Das Alter des Patienten könnte ein solches Attribut sein, es wurde abgesehen von der Expertenselektion von jedem anderen Verfahren als relevant erachtet.

5.2.2. k Means Clustering

Aufgrund der sehr ähnlichen Datensätze liegen die Clusterzentren immer sehr nahe beisammen und unterscheiden sich nur in wenigen Attributen wesentlich. Die deutlichsten Unterschiede treten beim Alter der Patienten, dem Geschlecht sowie der ICD10 Codierung auf. Dabei ist folgender Zusammenhang zu beobachten: Je älter ein Patient ist, desto kürzer ist seine prognostizierte Überlebenszeit. Zudem haben Männer eine schlechtere Prognose als Frauen. Die Lokalisation des Tumors scheint nur einen geringen Einfluss auf die Überlebenszeit zu haben.

Der K–Means Algorithmus wurde in insgesamt 11 unterschiedlichen Versuchskonstellationen genutzt. Diese unterschieden sich unter anderem in der Wahl des Merkmalsselektionsalgorithmus. Da alle diese Verfahren jeweils unterschiedliche Attributmengen als relevant erachtet hatten wurde in Tabelle 16 ausgewertet welches Attribut, wie häufig als relevant erachtet wurde. In diese Auswertung wurden der naive Ansatz der Merkmalsauswahl nicht mit einbezogen, da er alle Attribute als gleich wichtig erachtet.

Attributname	Vorkommen in %
UICC	90
ERSTE_R_KLASSIFIKATION	90
ANZAHL_METASTASEN	90
OP_INTENTION	60
PRIMAERTHERAPIE	60
ARZT_ANLASS	60
ANZAHL_TUMOREN	50
HISTO_GRADING	50
P_L	50
P_V	50
ICD10	40
ALTER	20
ERFASSUNGSANLASS	20
GESCHLECHT	20
BEHANDLUNGSANLASS	10

Tabelle 16: Häufigkeit der Verwendung der Attribute beim K–Means Clustering

Da sich in den meisten Fällen die Cluster in den Kaplan–Meier–Kurven zumindest ein wenig voneinander abhoben, kann davon ausgegangen werden, dass die häufiger verwendeten Attribute tatsächlich eine höhere Aussagekraft besitzen als die seltener verwendeten. Die Attribute sind in einer ähnlichen Reihenfolge wie beim naiven Bayes Klassifikator angeordnet (siehe Tabelle 17). Da die gleichen Merkmalsauswahlverfahren wie beim naiven Bayes Klassifikator verwendet wurden, war diese Ähnlichkeit zu erwarten.

Da bei den unterschiedlichen Versuchsdurchführungen die Anzahl der Cluster sich immer unterschieden, ist es schwer die Ergebnisse untereinander zu vergleichen. Es bleibt auf jeden Fall festzuhalten, dass im Vergleich zum naiven Ansatz nur die Merkmalsauswahl mittels Backward Elimination und Forward Selection besser von einander zu unterscheidende Kaplan–Meier–Schätzer lieferten. Dies ist vor allem unter dem Gesichtspunkt bemerkenswert, da die Kanten in den k-Diagrammen bei diesen beiden Verfahren am schwächsten ausgeprägt waren.
Insgesamt brachtet fällt auf, dass die k-Diagramme am glattesten (abgesehen von eventuell vorhandenen Kanten) sind, wenn die Anzahl der verwendeten Attribute hoch ist. Um möglichst deutlich hervorgehobene Kanten zu erhalten empfiehlt es sich folglich viele Attribute zu verwenden. Der Vergleich der k–Diagramme des naiven Ansatzes (Abbildung 7) unter Verwendung aller Attribute mit dem k–Diagramm Expertenselektion für zwei Attribute (Abbildung 13) zeigt diesen Unterschied deutlich.
Die Kaplan–Meier–Schätzer für die Expertenselektion und das Information Gain Verfahren zeichneten sich nicht so deutlich voneinander ab, wie beim naiven Ansatz. Während bei der Expertenselektion alle Cluster sich in zwei größere Gruppen aufteilen, liefert das Information Gain Verfahren im Gegensatz dazu zumindest Kurven, die sich untereinander deutlicher abheben. Es ist bei beiden Verfahren jedoch der Trend zu erkennen, dass mehr verwendete Attribute zu besseren Ergebnisse führen.

5.3. Datenqualität der Datensätze

Innerhalb dieser Arbeit wurde ausschließlich mit, im GTDS dokumentierten, Daten aus dem Tumorregister Heilbronn gearbeitet. Wie eingangs erwähnt, ist die Qualität der Daten ein wesentlicher Faktor für die Qualität des zu erwartenden Ergebnisses. Die Datenqualität wird laut der Deutsche Gesellschaft für Informations- und Datenqualität (DGIQ) in 15 Dimensionen bewertet. Daraus ergibt sich für die Datenqualität, der in dieser Arbeit behandelten Daten, folgende Bewertung:

Zugänglichkeit Aufgrund einer fehlenden pseudonymisierten Exportqualität und dem daher bedingten notwendigen Wissen über das Dokumentationssystem eher gering. Aufgrund der Sensibilität der Daten ist dies aber auch durchaus gewünscht.

Bearbeitkarkeit Nach dem Export der Daten konnten die Daten beliebig und auch automatisiert be– und verarbeitet werden.

Hohes Ansehen Durch das zuverlässige und seriöse Bild der SLK-Kliniken Heilbronn GmbH gegeben.

Fehlerfreiheit Kann ohne eine Referenz nicht umfassend beantwortet werden. Allerdings sind bei wenigen Datensätzen vereinzelt Fälle mit unmöglichen Wertekonstellationen entdeckt worden.

Objektivität Aufgrund vieler verwendeter objektiver Klassifikations–Skalen und eindeutig zu erfassender Werte bei gleichzeitig verhältnismäßig geringem Anteil von Freitextfelder, erscheint die Objektivität dieses Datenbestandes sehr hoch.

Glaubwürdigkeit Da die Daten ausschließlich von Experten erfasst und dokumentiert werden, sind die Daten in höchstem Maße glaubwürdig.

Eindeutige Auslegbarkeit In vielen Fällen werden Attribute codiert dokumentiert. über den Schlüssel ist dadurch eine eindeutiges Verständnis möglich.

Einheitliche Darstellung Aufgrund der hohen Standardisierung natürlicherweise sehr hoch.

Übersichtlichkeit Ist nur bedingt gegeben. Eine Vielzahl an Attributen erschwert den Einstieg. Allerdings wurde im Rahmen der Thesis nur mit dem csv-Export der Daten gearbeitet. Die Präsentation der Daten kann innerhalb des Dokumentationssystems besser sein.

Verständlichkeit Aufgrund der guten Online-Dokumentation der Tabellen gegeben.

Relevanz Für diese Arbeit waren lediglich maximal 18 der 195 Attribute interessant. Allerdings, war der Hauptzweck der Dokumentation auch nicht das Ausarbeiten dieser Thesis, was diesen Overhead verständlich macht.

Angemessener Umfang An manchen Stellen ist für die Ausarbeitung einer solchen wissenschaft-

lichen Arbeit die Dokumentation etwas knapp gehalten.

Vollständigkeit Viele Attribute sind nur selten oder nie dokumentiert (siehe Anhang A Tabelle 17)

Wertschöpfung Es konnten, wie in dieser Arbeit gezeigt, Ergebnisse auf der Grundlage dieser Daten erzielt werden.

Aktualität Leider ohne die Möglichkeit einer Pseudonymisierungsschittstelle an das Dokumentationssystem nicht gegeben.

Insgesamt ist also die Datenqualität für diese Arbeit ausreichend gewesen, das einzige echte Manko waren die teilweise vielen fehlenden Werte bei einzelnen Attributen. Viele der anderen Probleme könnten sich in Zukunft durch die Implementierung einer Pseudonymisierungsschnittstelle lösen.

6. Fazit

In dieser Arbeit wurden der naive Bayes Klassifikator gegen das K–Means Clustering bezüglich ihrer Prognosefähigkeit der Überlebenszeit von Tumorpatienten anhand des initial dokumentierten allgemeinen Zustandes berechnet. Dabei wurden neben der eigentlichen Data Mining–Verfahren auch verschiedene Merkmalsauswahlverfahren gegeneinander getestet.
Unter den Merkmalsauswahlverfahren erwieß sich die Backward Elimination, bei der vorhandenen Konstellation der Menge und Verteilung der Datensätze, als das beste der hier getesteten Verfahren. Dies ist zum einen aus den errechneten Fehlermaßen der naiven Bayes Klassifizierung sowie der Unterscheidbarkeit der Kaplan–Meier–Kurven des K–Means Clusterings abzulesen. Als zweitbestes Merkmalsauswahlverfahren erwies sich das Information Gain–Verfahren, jedoch nur sofern eine höhere Anzahl an Attributen (7,8 oder 10) verwendet wurde. Die Ergebnisse der Forward Selection waren nach dem K–Means Clustering deutlich besser als beim naiven Vergleichsansatz. Nach einer Klassifikation mit Hilfe des naives Bayes–Verfahrens lieferte die Forward Selection allerdings als schlechtere Resultate als der naive Ansatz. Dies zeigt, dass je nach gewältem Data–Mining–Verfahren eine andere Merkmalauswahl stattfinden muss.
Die Expertenselektion als Merkmalsauswahlverfahren schnitt in allen Versuchsreihen unerwartet schlecht ab. Dies zeigt die Komplexität der MErkmalsauswahl, die es selbst für Experten schwer macht eine valide Aussage bezüglich der Relevanz von Attributen zu treffen.

Bei der Auswahl der Attribute fällt auf, dass die konkrete Auswahl zwar vom jeweiligen Data–Mining–Verfahren abhängig ist, es jedoch trotzdem Ähnlichkeiten in Häufigkeit der Verwendung der einzelnen Attribute gibt. So erwiesen sich das UICC–Staging, die erste R–Klassifikation und die Anzahl der Metastasen des Tumors als sehr ausgekräftig. Diese Attribute waren sowohl für den naiven Bayes Klassifikator als auch für das K–Means Clustering die drei am häufigsten verwendeten Attribute.

Aufgrund der Möglichkeit der patientenindividuellen Prognose eines Überlebenszeit-Wahrscheinlichkeit-Histogrammes erwies sich der naive Bayes Klassifikator im Laufe der Thesis als intuitiver und besser nachvollziehbar.
Dies ist auch durch die nahe beisammen liegenden Clusterzentren des K–Means Clusterings zu erklären. Diese unterschieden sich alle häufig nur in der Ausprägung eines einzelnen Attributes, wodurch die Zuordnung eines Patienten zu einem bestimmten Clusterzentrum für den Betrachter nicht jedes mal nachvollziehbar erscheinen mag. Die so erzeugten Kaplan–Meier–Schätzer hoben sich jedoch zum Teil deutlich voneinander ab, was die schwer nachvollziehbare Lage der Clusterzentren bestätigt.

6.1. Ausblick

Diese Arbeit kann den Beginn einer Serie von weiteren Bachelorarbeiten in Kooperation mit dem Tumorzentrum Heilbronn–Franken darstellen. Zum einen könnten aus dieser Kooperation komplett neue Arbeiten hervorgehen, zum anderen könnte auf der Basis dieser Thesis weitergearbeitet werden.
Auf Grundlage dieser Arbeit könnten weitere Data Mining–Verfahren in Betracht gezogen werden, um diese untereinander und mit den hier vorgestellten Verfahren zu vergleichen. Auch

könnte versucht werden die Präsentation des hier vorgestellten naiven Bayes Klassifikators für Ärzte zu erhöhen. Dazu müsste aus den Überlebenszeit–Histogrammen des naiven Bayes Klassifikators Kaplan–Meier–Schätzer errechnet werden. Eine Möglichkeit dies zu tun wäre, zufällig aus den diskretisierten Überlebenszeitintervallen wieder konkrete Werte zu generieren. Da beim naiven Bayes Klassifikator in der Trainingsphase ausschließlich Tumortode berücksichtigt wurden, würde man dann für jeden einzelnen Patienten ein Kaplan–Meier–Diagramm ohne Zensuren erhalten.
Des Weiteren könnte die Ergebnisse auf Basis von nur knapp 4000 Datensätzen durch Hinzunahme eines größeren Tumorregisters validiert werden. Zudem könnte mit den gleichen Prozessen auch andere Tumorlokalisationen untersucht werden.
Auch könnte der Schritt weg von der Grundlagenfoschung gewagt werden, um mit Hilfe der hier präsentierten Ergebnisse eine Überlebenszeit–Prognose–Anwendung für Onkologen zu entwickeln, welches aus dem initialen Status eines Tumorpatienten eine Prognose auf dessen Überlebenszeit errechnet.

Abkürzungsverzeichnis

CRISP-DM Cross–Industry Standard Process for Data Mining

csv Comma Separated Value

DGIQ Deutsche Gesellschaft für Informations- und Datenqualität

GTDS Gießener Tumordokumentationssystem

ICD International Statistical Classification of Diseases and Related Health Problems

TNM Tumor Nodes Metastasis

UICC Union Internationale Contra le Cancer

Literatur

[AE07] Arihito Endo, Takeo Shibata und Hiroshi T.: Comparison of Seven Algorithms to predict Breast Cancer Survival. In: *Biomedical Soft Computing and Human Sciences* (2007)

[al.07] al., Xindong W.: Top 10 algorithms in data mining. In: *Knowledge and Information Systems* (2007)

[Ber13] Berthold, Michael: *Onlinepräsenz des Konstanz Information Miner*. `http://www.knime.org/`, November 2013

[Bor10] Borgelt, Christian: Induktion von Entscheidungsbäumen. (2010)

[DUA13] Dr.med. Udo Altmann, Frank R. Katz und Alexander Q.: *Webauftritt des Gießener Tumordokumentationssystem*. `http://www.med.uni-giessen.de/akkk/gtds/`, November 2013

[ecl13] *Eclipse Indigo*. `http://www.eclipse.org/indigo/`, November 2013

[Efr88] Efron, Bradley: Logistic Regression, Survival Analysis and the Kaplan-Meier Curve. In: *American Statistical Association* (1988)

[GHJ94] George H. John, Ron Kohavi und Karl P.: Irrelevant Features and the Subset Selection Problem. In: *Machine Learning: Proceedings of the Eleventh International Conference* (1994)

[gra14] *Grading*. `http://www.darmkrebs.de/frueherkennung-diagnose/stadieneinteilung/staging-grading/`, Januar 2014

[IHW11] Ian H. Witten, Eibe Frank und Mark A. H.: *Data Mining*. Morgan Kaufmann Publishers (Elsevier), 2011

[IM13] Ingo Mierswa, Andy Menzies und Andrew D.: *Webauftritt von Rapid Miner*. `http://rapidminer.com/`, November 2013

[Mac67] MacQueen, J. B.: *Some Methods for classification and Analysis of Multivariate Observations*. University of California Press : Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, 1967

[Mar14] Martens, Uwe: *Medizinische Klinik III des Klinikums am Gesundbrunnen Heilbronn*. `http://www.slk-kliniken.de/Innere-Medizin-III-Haematolog.50.0.html`, Januar 2014

[NM99] Nicole Menche, Arne S.: *Mensch Körper Krankheit*. München : Urban & Fischer, 1999

[Nor12] North, Matthew: *Data Mining for the Masses.* Global Text Project Book, 2012

[OM09] Oscar Marban, Gonzalo Mariscal und Javier S.: A Data Mining and Knowledge Discovery Process Model. In: *Data Mining and Knowledge Discovery in Real Life Applications* (2009)

[Pap06] Papathanassiou, Drosoula-Aliki: *Prognosefaktoren des papillären Schilddrüsenkarzinoms.* Heinrich-Heine-Universität Düsseldorf, 2006

[PNT06] Pang-Ning Tan, Michael Steinbach und Vipin K.: *Introduction to Data Mining.* Pearson International Edition, 2006

[r13] *The R Project for Statistical Computing.* `http://www.r-project.org/`, November 2013

[RF11] Roland Fuchs, Ulf Neumann und Christian T. Dorothee Guggenberger G. Dorothee Guggenberger: *GI-Tumore.* Stolberg : Nora-Verlag, 2011

[RM08] Rainer Münz, Steffen K.: Sterblichkeit und Todesursachen. In: *Berlin-Institut für Bevölkerung und Entwicklung* (2008)

[Sek13] Sekulla, C.: *Onlinepräsenz des Kaplan-Meier-Überlebensstatistik Excel-Makros des Universitätsklinikums Halle (Saale).* `https://www.medizin.uni-halle.de/index.php?id=1358`, Dezember 2013

[SG11] Shelly Gupta, Dharminder Kumar und Anand S.: Data Mining Classification techniques applied for breast cancer diagnosis and prognosis. In: *Indian Journal of Computer Sciene and Engineering* (2011)

[slk14] *Onlinepräsenz der SLK-Kliniken Heilbronn.* `http://www.slk-kliniken.de/`, Februar 2014

[Tih14] Tihanyi, Balazs: *Onlinepräsenz von NClass.* `http://sourceforge.net/projects/nclass/`, Februar 2014

[tnm14] *Erläuterung TNM-Klassifikation.* `http://www.uicc.org/resources/tnm`, Februar 2014

A. Anhang

A.1. Attribute der exportierten Datensätze

Tabelle 17 zeigt alle von GTDS im Export der Auswertungstabelle vorhandenen Attribute. Dazu wird dargestellt, ob das jeweilige Attribut für das Data Mining direkt genutzt wurde. Ob dies der Fall ist, ist in der Spalte „Aufgenommen“ vermerkt. In der Spalte „Begründung“ wird erörtert, warum das jeweilige Attribut nicht direkt in die Attributmenge zur Auswertung übernommen wurde. Attribute, die zur Berechnung anderen Werte notwendig waren, später aber nicht im Attributsatz zu finden sind, wurden ebenfalls mit „NEIN“ gekennzeichnet. Dazu zählen beispielsweise alle dokumentierten Zeitpunkte, die zur Berechnung den Zeiträumen, wie dem Behandlungszeitraum oder dem Alter des Patienten notwendig waren, aber in der späteren Auswertung nicht auftauchen. Ebenfalls betrifft dies das TNM-Staging, welches in der Vorverarbeitung, siehe Abschnitt 3.3, zur Berechnung des UICC-Stagings benötigt wurde.

Attributname	Typ	fehlende Werte	Aufgenommen	Begründung
ABTEILUNG1	Ganzzahl	0	NEIN	GTDS-Interne Nummer
ABTEILUNG2	Ganzzahl	0	NEIN	GTDS-Interne Nummer
ABTEILUNG3	Ganzzahl	686	NEIN	GTDS-Interne Nummer
ALLE_ABTEILUNGEN	Fließkommazahl	0	NEIN	GTDS-Interne Nummer
ALLE_AERZTE	Fließkommazahl	328	NEIN	GTDS-Interne Nummer
ALLE_BEGLEITERKRANKUNGEN	Freitext	3757	NEIN	zu viele fehlende Werte
ALLE_BESTRAHLUNGEN	Freitext	3163	NEIN	initial nicht vorhanden
ALLE_FOLGEERKRANKUNGEN	Freitext	3697	NEIN	zu viele fehlende Werte
ALLE_INNEREN	Freitext	2523	NEIN	zu komplex
ALLE_METASTASEN	Freitext	2581	NEIN	zu komplex
ALLE_OPERATIONEN	Freitext	961	NEIN	zu komplex
ALLE_VORERKRANKUNGEN	Freitext	3743	NEIN	zu komplex
ANN_ARBOR_ALLGEMEIN	Freitext	3759	NEIN	zu viele fehlende Werte
ANN_ARBOR_DATUM	Freitext	3758	NEIN	zu viele fehlende Werte
ANN_ARBOR_EXTRA	Freitext	3759	NEIN	zu viele fehlende Werte
ANN_ARBOR_HERKUNFT	Freitext	3758	NEIN	zu viele fehlende Werte
ANN_ARBOR_LFDNR	Freitext	3758	NEIN	zu viele fehlende Werte
ANN_ARBOR_STADIUM	Freitext	3758	NEIN	zu viele fehlende Werte
ANZAHL_ABTEILUNGEN	Ganzzahl	0	NEIN	initial nicht vorhanden
ANZAHL_AERZTE	Ganzzahl	0	NEIN	initial nicht vorhanden
ANZAHL_AUFENTHALTE	Freitext	3429	NEIN	zu viele fehlende Werte
ANZAHL_BESTRAHLUNGEN	Ganzzahl	0	NEIN	Behandlungsverlauf
ANZAHL_FOLGEERKRANKUNGEN	Ganzzahl	0	NEIN	initial nicht vorhanden
ANZAHL_HISTOLOGIEN	Ganzzahl	0	NEIN	Behandlungsverlauf
ANZAHL_INNERE	Ganzzahl	0	NEIN	Behandlungsverlauf
ANZAHL_METASTASEN	Ganzzahl	0	JA	
ANZAHL_NACHSORGEN	Ganzzahl	0	NEIN	Behandlungsverlauf
ANZAHL_OPERATIONEN	Ganzzahl	0	NEIN	Behandlungsverlauf
ANZAHL_R_KLASSIFIKATIONEN	Ganzzahl	0	NEIN	irrelevant
ANZAHL_SONSTIGE	Ganzzahl	0	NEIN	Behandlungsverlauf
ANZAHL_TEILBESTRAHLUNGEN	Ganzzahl	3163	NEIN	zu viele fehlende Werte
ANZAHL_TUMOREN	Ganzzahl	0	JA	
ANZAHL_ZIELGEBIETE	Ganzzahl	3166	NEIN	zu viele fehlende Werte
APPLIKATIONSART	Freitext	3553	NEIN	zu viele fehlende Werte
APPLIKATIONSTECHNIK	Freitext	3759	NEIN	zu viele fehlende Werte
ARZT_ANLASS	Freitext	173	JA	
ARZT1	Ganzzahl	2084	NEIN	GTDS-Interne Nummer
AUFNAHMEDATUM	Datum	157	NEIN	Zeitpunkt irrelevant

AUTOPSIE	Freitext	2063	NEIN	irrelevant
BEGINN_NACHSORGE	Datum	2365	NEIN	Zeitpunkt irrelevant
BEHANDLUNGSANLASS	Freitext	3023	JA	
BENUTZER	Freitext	0	NEIN	irelevant
BESTRAHLUNG_ABTEILUNG	Ganzzahl	3163	NEIN	GTDS-Interne Nummer
BESTRAHLUNG_DATUM	Datum	3165	NEIN	Zeitpunkt irrelevant
BESTRAHLUNG_DF_ABT_ID	Ganzzahl	3163	NEIN	GTDS-Interne Nummer
BESTRAHLUNG_DF_ARZT_ID	Freitext	3759	NEIN	GTDS-Interne Nummer
BESTRAHLUNG_FREITEXT	Freitext	3163	NEIN	zu komplex
BESTRAHLUNG_NUMMER	Ganzzahl	3163	NEIN	GTDS-Interne Nummer
C_STADIUM	Freitext	3210	NEIN	durch UICC ersetzt
DATUM_DER_AUSWERTUNG	Datum		NEIN	nicht exportiert
DATUM_ERSTE_METASTASE	Datum	2606	NEIN	Zeitpunkt irrelevant
DATUM_ERSTE_PROGRESSION	Datum	2775	NEIN	Zeitpunkt irrelevant
DATUM_ERSTES_REZIDIV	Datum	3093	NEIN	Zeitpunkt irrelevant
DEFIN_LOKALE_RADIKALITAET	Freitext	1380	NEIN	Infos über Behandlungsverlauf
DIAGNOSE_ABTEILUNG	Ganzzahl	0	NEIN	GTDS-Interne Nummer
DIAGNOSEDATUM	Datum	82	NEIN	Zeitpunkt irrelevant
DIAGNOSE_DF_ABT_ID	Ganzzahl	28	NEIN	GTDS-Interne Nummer
DIAGNOSE_DF_ARZT_ID	Ganzzahl	3160	NEIN	GTDS-Interne Nummer
DIAGNOSETEXT	Freitext	0	NEIN	zu komplex
DIAGSICH_HOECHSTE	Freitext	12	NEIN	zu komplex
ERFASSUNGSANLASS	Freitext	2993	JA	
ERSTE_LOKALE_RADIKALITAET	Freitext	1380	NEIN	nicht verstanden
ERSTE_R_KLASSIFIKATION	Freitext	1507	JA	
ERSTES_LOK_REZIDIVART	Freitext	3502	NEIN	zu viele fehlende Werte
ERSTES_LOK_REZIDIVDATUM	Datum	3502	NEIN	Zeitpunkt irrelevant
ERSTES_LOK_REZIDIVVERLAUF	Ganzzahl	3502	NEIN	zu viele fehlende Werte
ERSTES_REZIDIV	Freitext	3091	NEIN	zu viele fehlende Werte
FOLGEERKRANKUNG1	Freitext	3697	NEIN	zu viele fehlende Werte
FOLGEERKRANKUNG2	Freitext	3751	NEIN	zu viele fehlende Werte
GEBURTSDATUM	Datum	0	NEIN	Zeitpunkt irrelevant
GESAMTDAUER_AUFENTHALTE	Freitext	3647	NEIN	zu viele fehlende Werte
GESAMTDOSIS	Freitext	3610	NEIN	zu viele fehlende Werte
GESCHLECHT	Binärwert	2	JA	
GY_GBQ	Freitext	3544	NEIN	
HAUSARZT	Freitext		NEIN	nicht exportiert
HISTO_AUFLAGE	Freitext	10	NEIN	irrelevant
HISTO_DATUM	Datum	196	NEIN	Zeitpunkt irrelevant
HISTO_DIAGNOSE	Binärwert	102	NEIN	nicht aussagekräftig
HISTO_GRADING	Freitext	561	JA	
HISTO_HAUPT_NEBEN	Binärwert	18	NEIN	nicht aussagekräftig
HISTO_HERKUNFT	Binärwert	12	NEIN	nicht aussagekräftig
HISTO_LFDNR	Ganzzahl	7	NEIN	GTDS-Interne Nummer
HISTOLOGIE	Ganzzahl	8	NEIN	zu komplex
HISTOLOGIE2	Freitext	3759	NEIN	zu viele fehlende Werte
HISTO_SICHERUNGSDATUM	Datum	55	NEIN	Zeitpunkt irrelevant
ICD10	Freitext	0	JA	
ICD9	Freitext	3759	NEIN	zu viele fehlende Werte
INNERE_ABTEILUNG	Ganzzahl	2523	NEIN	GTDS-Interne Nummer
INNERE_ANZAHL_ZYKLEN	Ganzzahl	2523	NEIN	nicht aussagekräftig
INNERE_BEGINN	Datum	2525	NEIN	Zeitpunkt irrelevant
INNERE_DF_ABT_ID	Ganzzahl	2554	NEIN	GTDS-Interne Nummer
INNERE_DF_ARZT_ID	Ganzzahl	3516	NEIN	GTDS-Interne Nummer
INNERE_FREITEXT	Freitext	2523	NEIN	zu komplex
INNERE_NUMMER	Ganzzahl	2523	NEIN	GTDS-Interne Nummer
INNERE_PROTOKOLL_ID	Ganzzahl	3205	NEIN	GTDS-Interne Nummer
INNERE_PROTOKOLL_TYP	Freitext	2526	NEIN	zu komplex

KKR_EINWILLIGUNG	Freitext	3759	NEIN	zu viele fehlende Werte
KLIN_L	Freitext	3758	NEIN	zu viele fehlende Werte
KLIN_M	Freitext	3729	NEIN	nicht verstanden
KLIN_MET	Freitext	2949	NEIN	durch UICC ersetzt
KLIN_N	Freitext	2956	NEIN	durch UICC ersetzt
KLIN_P_M	Freitext	2972	NEIN	irrelevant
KLIN_P_N	Freitext	2979	NEIN	irrelevant
KLIN_PNI	Freitext	3759	NEIN	zu viele fehlende Werte
KLIN_P_T	Freitext	2978	NEIN	irrelevant
KLIN_S	Freitext	3759	NEIN	zu viele fehlende Werte
KLIN_T	Freitext	2956	NEIN	durch UICC ersetzt
KLIN_TNM_AUFLAGE	Ganzzahl	2959	NEIN	irrelevant
KLIN_TNM_AUSWERTUNGS_RELEVANT	Freitext	3758	NEIN	zu viele fehlende Werte
KLIN_TNM_DATUM	Datum	2945	NEIN	Zeitpunkt irrelevant
KLIN_TNM_HERKUNFT	Binärwert	2942	NEIN	irrelevant
KLIN_TNM_LFDNR	Ganzzahl	2942	NEIN	GTDS-Interne Nummer
KLIN_V	Freitext	3759	NEIN	zu viele fehlende Werte
KLIN_Y_SYMBOL	Freitext	3754	NEIN	zu viele fehlende Werte
LETZTE_INFO_DATENART	Freitext	0	NEIN	nach Berechnungen löschen
LETZTE_INFO_DATUM	Datum	64	NEIN	Zeitpunkt irrelevant
LETZTE_INFO_LFDNR	Ganzzahl	0	NEIN	GTDS-Interne Nummer
LETZTE_NA_OHNE_PROGRESSION	Datum	2885	NEIN	Zeitpunkt irrelevant
LETZTER_ABSCHLUSS_DATUM	Datum	2665	NEIN	Zeitpunkt irrelevant
LETZTER_ABSCHLUSS_GRUND	Freitext	2665	NEIN	nach Berechnungen löschen
LETZTER_ABSCHLUSS_LFDNR	Ganzzahl	543	NEIN	GTDS-Interne Nummer
LETZTE_R_KLASSIFIKATION	Freitext	1606	NEIN	nur Initialzustand interessant
LETZTER_STATUS	Freitext	543	NEIN	zu komplex
LETZTER_STATUS_DATENART	Binärwert	543	NEIN	GTDS-Interne Nummer
LETZTER_STATUS_DATUM	Datum	544	NEIN	Zeitpunkt irrelevant
LETZTE_STATUS_LFDNR	Ganzzahl	2665	NEIN	GTDS-Interne Nummer
LOKALISATION	Ganzzahl	1	NEIN	redundant zu ICD10
LOKALISATION2	Freitext	3759	NEIN	zu viele fehlende Werte
LOK_AUFLAGE	Ganzzahl	2	NEIN	irrelevant
LOK_HAUPT_NEBEN	Binärwert	1	NEIN	nicht aussagekräftig
LOK_SEITE	Freitext	3	NEIN	nicht aussagekräftig
METASTASE1	Freitext	2581	NEIN	zu komplex
METASTASE2	Freitext	3217	NEIN	zu komplex
NACHFRAGEARZT	Ganzzahl	328	NEIN	GTDS-Interne Nummer
NAME	Freitext		NEIN	nicht exportiert
OP_ABTEILUNG	Ganzzahl	888	NEIN	GTDS-Interne Nummer
OP_BEZEICHNUNG	Freitext	941	NEIN	zu komplex
OP_DATUM	Datum	941	NEIN	Zeitpunkt irrelevant
OP_DF_ABT_ID	Ganzzahl	895	NEIN	GTDS-Interne Nummer
OP_DF_ARZT_ID	Freitext	3752	NEIN	GTDS-Interne Nummer
OP_DURCHGEFUEHRT	Freitext	888	NEIN	zu komlex
OP_INTENTION	Freitext	1761	JA	
OP_NUMMER	Ganzzahl	888	NEIN	GTDS-Interne Nummer
OP_SCHLUESSEL	Freitext	1883	NEIN	irrelevant
OP_SCHLUESSEL_AUFLAGE	Freitext	970	NEIN	irrelevant
ORTSKENNZAHL	Ganzzahl	47	NEIN	irrelevant
PAT_ID	Ganzzahl	0	JA	
P_L	Ganzzahl	2534	JA	
PLZ	Ganzzahl	6	NEIN	irrelevant
PLZ_BEI_DIAGNOSE	Ganzzahl	6	NEIN	irrelevant
P_M	Freitext	3630	NEIN	nicht verstanden
P_MET	Freitext	809	NEIN	durch UICC ersetzt
P_N	Freitext	719	NEIN	durch UICC ersetzt
P_P_M	Freitext	869	NEIN	irrelevant

P_P_N	Freitext	723	NEIN	irrelevant
P_PNI	Ganzzahl	3653	NEIN	zu viele fehlende Werte
P_P_T	Binärwert	715	NEIN	irrelevant
PRIMAERTHERAPIE	Freitext	923	JA	
PRIMFALL	Freitext	3759	NEIN	zu viele fehlende Werte
P_S	Freitext	3758	NEIN	zu viele fehlende Werte
P_STADIUM	Freitext	2199	NEIN	durch UICC ersetzt
P_T	Freitext	715	NEIN	durch UICC ersetzt
P_TNM_AUFLAGE	Ganzzahl	787	NEIN	irrelevant
P_TNM_AUSWERTUNGS_RELEVANT	Freitext	3758	NEIN	zu viele fehlende Werte
P_TNM_DATUM	Datum	765	NEIN	Zeitpunkt irrelevant
P_TNM_HERKUNFT	Binärwert	715	NEIN	zu viele fehlende Werte
P_TNM_LFDNR	Ganzzahl	715	NEIN	GTDS-Interne Nummer
P_V	Ganzzahl	2573	JA	
P_Y_SYMBOL	Freitext	3524	NEIN	zu viele fehlende Werte
REFERENZ	Freitext	3759	NEIN	GTDS-Interne Nummer
REGISTER	Binärwert	0	NEIN	irrelevant
SATZNUMMER	Ganzzahl	0	NEIN	irrelevamt
SONSTIGE_DATUM	Freitext	3759	NEIN	zu viele fehlende Werte
SONSTIGE_HERKUNFT	Freitext	3748	NEIN	irrelevant
SONSTIGE_ID	Freitext	3748	NEIN	GTDS-Interne Nummer
SONSTIGE_KUERZEL	Freitext	3748	NEIN	zu viele fehlende Werte
SONSTIGE_LFDNR	Freitext	3748	NEIN	zu viele fehlende Werte
SONSTIGE_NAME	Freitext	3748	NEIN	zu viele fehlende Werte
SONSTIGE_STADIUM	Freitext	3752	NEIN	zu viele fehlende Werte
SONSTIGETHERAPIE	Freitext	3270	NEIN	nicht aussagekräftig
STERBEDATUM	Datum	1934	NEIN	Zeitpunkt irrelevant
STERBEDATUM_EXAKT	Binärwert	1944	NEIN	
STRAHLENART	Freitext	3759	NEIN	zu viele fehlende Werte
TEIL_BESTR_BEGINN	Datum	3168	NEIN	Zeitpunkt irrelevant
THERAPIEEBEGINN	Datum	925	NEIN	Zeitpunkt irrelevant
THERAPIEENDE	Datum	930	NEIN	Zeitpunkt irrelevant
TRANSFORMATION_DATUM	Freitext	3759	NEIN	zu viele fehlende Werte
TRANSFORMATION_HISTO_AUFLAGE	Freitext	3759	NEIN	zu viele fehlende Werte
TRANSFORMATION_HISTO_CODE	Freitext	3759	NEIN	zu viele fehlende Werte
TRANSFORMATION_VERLAUF	Freitext	3759	NEIN	zu viele fehlende Werte
TUMORFOLGENUMMER	Freitext	3759	NEIN	zu viele fehlende Werte
TUMORFREIHEIT_VERLAUF	Ganzzahl	1661	NEIN	GTDS-Interne Nummer
TUMOR_ID	Ganzzahl	0	NEIN	GTDS-Interne Nummer
TUMORTOD	Freitext	1959	NEIN	durch UICC ersetzt
VORGANG_ID	Ganzzahl		NEIN	GTDS-Interne Nummer
VORNAME	Freitext		NEIN	nicht exportiert
ZEITPUNKT_TUMORFREIHEIT	Datum	1661	NEIN	Zeitpunkt irrelevant
ZENTKENN	Ganzzahl	2492	NEIN	irrelevant
ZIELGEBIET1	Freitext	3756	NEIN	zu viele fehlende Werte
ZIELGEBIET2	Freitext	3759	NEIN	zu viele fehlende Werte

Tabelle 17: Attribute der exportierten Datensätze

A.2. Attribute für das Data Mining

Tabelle 18 zeigt eine Übersicht über die Attribute, die für das Data Mining verwendet wurden, inklusive der Vorverarbeitungsschritte, die an dem jeweiligen Attribut vorgenommen wurde.

Attributname	Typ	fehlende Werte	Vorverarbeitung
PAT_ID	Ganzzahl	0	keine
ANZAHL_TUMOREN	Fließkommazahl	0	normiert
GESCHLECHT	Binärzahl	0	fehlende Werte aufgefüllt
ICD10	Freitext	0	pro Ausprägung ein boolsches Attribut
HISTO_GRADING	Freitext	0	zusätzlichen Wert für nicht angegebe Werte eingefügt, pro Ausprägung ein boolsches Attribut
PRIMAERTHERAPIE	Freitext	0	Primärtheraphie aus Daten ermitteln, pro Ausprägung ein boolsches Attribut
ERSTE_R_KLASSIFIKATION	Fließkommazahl	0	zusätzlichen Wert für nicht angegebe Werte eingefügt, pro Ausprägung ein boolsches Attribut
ANZAHL_METASTASEN	Fließkommazahl	0	normiert
OP_INTENTION	Freitext	0	pro Ausprägung ein boolsches Attribut
P_L	Fließkommazahl	0	fehlende Werte mit 0 auffüllen, normieren
P_V	Fließkommazahl	0	fehlende Werte mit 0 auffüllen, normieren
BEHANDLUNGSANLASS	Freitext	0	zusätzlichen Wert für nicht angegebe Werte eingefügt, pro Ausprägung ein boolsches Attribut
ERFASSUNGSANLASS	Freitext	0	zusätzlichen Wert für nicht angegebe Werte eingefügt, pro Ausprägung ein boolsches Attribut
ARZT_ANLASS	Freitext	0	pro Ausprägung ein boolsches Attribut
ALTER	Fließkommazahl	0	aus Geburtsdautm und Diagnosedatum berechnet, dann normiert
UICC	Fließkommazahl	405	aus TNM-Staging berechnet, dann normiert
TIMES	Ganzzahl	0	Aus Überlebensdauer berechnet
EVENTS	Ganzzahl	0	Gibt n ob Tod durch Tumor oder Zensur

Tabelle 18: Attribute für das Data Mining

A.3. Ausgewählte Attribute der Merkmalsauswahlverfahren

Auf den folgenden drei Seiten sind die Atribute tabellarisch aufgelistet, welche von den einzelnen Merkmalsauswahlverfahren als relevant erachtet wurden. Tabelle 19 stellt die Auswahl für das K–Means Clustering da. Aus Platzgründen wurden folgende Abkürzungen verwendet:

NS: No Selection

BE: Backward Elimination

FS: Forward Selection

IG: Information Gain

ES: Expert Selection

Die Zahlen hinter IG und ES stehen für die Anzahl der Attribute auf welche das jeweilige Verfahren beschränkt wurde. Ein „x“ in einem Feld der Tabelle bedeutet, dass das jeweilige Attribut von dem entsprechenden Merkmalsauswahlverfahren als relevant angesehen wurde.

Die Tabellen 20 und 21 stellen die Merkmalsauswahl für das naive Bayes Klassifikationsverfahren nach äquidistanter und äquifrequenter Diskretisierung der Überlebenszeit dar. Dabei wurde in jedem Feld vermerkt bei welcher Anzahl an Diskretisierungsintervallen (2,3,4,5 oder 10) das jeweilige Attribut als relevant erachtet wurde. Bei Felder die mit einem „x“ gekennzeichnet wurden, wurde das entsprechende Attribut bei allen Anzahlen der Diskretisierungsintervallen genutzt.

	NS	ES2	ES7	ES8	ES10	IG2	IG7	IG8	IG10	FS	BE
ANZAHL_TUMOREN	x	x	x	x	x						x
GESCHLECHT	x									x	x
ICD10	x			x	x					x	x
HISTO_GRADING	x		x	x	x					x	x
PRIMAERTHERAPIE	x				x		x	x	x	x	x
ERSTE_R_KLASSIFIKATION	x		x	x	x		x	x	x	x	x
ANZAHL_METASTASEN	x	x	x	x	x	x	x	x	x		x
OP_INTENTION	x				x		x	x	x	x	
P_L	x		x	x	x					x	x
P_V	x		x	x	x					x	x
BEHANDLUNGSANLASS	x									x	x
ERFASSUNGSANLASS	x									x	x
ARZT_ANLASS	x						x	x	x	x	x
ALTER	x									x	x
UICC	x		x	x	x	x	x	x	x	x	x

Tabelle 19: Merkmalsauswahl beim K–Means Clustering

	NS	ES2	ES7	ES8	ES10	IG2	IG7	IG8	IG10	FS	BE
ANZAHL_TUMOREN	x	x	x	x	x		x	x	x		
GESCHLECHT	x									x	x
ICD10	x			x	x			x	x		4
HISTO_GRADING	x		x	x	x						2,5,10
PRIMAERTHERAPIE	x				x	x	x	x	x		5,10
ERSTE_R_KLASSIFIKATION	x		x	x	x		x	x	x		
ANZAHL_METASTASEN	x	x	x	x	x		x	x	x		2,4
OP_INTENTION	x				x		x	x	x		3
P_L	x		x	x	x				x		
P_V	x		x	x	x						
BEHANDLUNGSANLASS	x										3,4,10
ERFASSUNGSANLASS	x										2,3
ARZT_ANLASS	x						x	x	x		2,4,5
ALTER	x					x	x	x	x		x
UICC	x		x	x	x				x		

Tabelle 20: Merkmalsauswahl beim naiven Bayes Klassifikator mit äquidistant disrketisierter Überlebenszeit

	NS	ES2	ES7	ES8	ES10	IG2	IG7	IG8	IG10	FS	BE
ANZAHL_TUMOREN	x	x	x	x	x						2,3,5,10
GESCHLECHT	x										2,3,5
ICD10	x			x	x		x	x	x		
HISTO_GRADING	x		x	x	x				x	4,5,10	x
PRIMAERTHERAPIE	x				x			x	x	2,3,4,5	2,3,5,10
ERSTE_R_KLASSIFIKATION	x		x	x	x	x	x	x	x	2,3,5	x
ANZAHL_METASTASEN	x	x	x	x	x				x	2,4	2,3,4,10
OP_INTENTION	x				x		x	x	x	2,4	2,3,4,10
P_L	x		x	x	x						3,5
P_V	x		x	x	x						
BEHANDLUNGSANLASS	x						x	x	x	5	x
ERFASSUNGSANLASS	x						x	x	x	4,5	x
ARZT_ANLASS	x						x	x	x	2,3,5	x
ALTER	x									2,4	x
UICC	x		x	x	x	x	x	x	x	x	x

Tabelle 21: Merkmalsauswahl beim naiven Bayes Klassifikator mit äquifrequent disrketisierter Überlebenszeit

A.4. Verteilung Äquifrequente Diskretisierung

Alle Werte sind auf eine Nachkommastellen gerundet. Dass bei der bei einer äquifrequenten Diskretisierung nicht alle Diskretisierungsintervalle den gleichen Anteil an Elementen enthalten, liegt daran, dass im niedrigen Wertebereich viele Patienten auf wenige Werte verteilt werden, dadurch ist eine exakt homogene Verteilung nicht möglich.

Intervall (Überlebenszeit in Tagen)	**Anteil in %**
$[-\infty; 199, 5]$	$28, 2$
$[199, 5; 1025, 5]$	$40, 6$
$[1025, 5; \infty]$	$31, 2$

Tabelle 22: Äquifrequente Diskretisierung in 3 Intervalle

Intervall (Überlebenszeit in Tagen)	**Anteil in %**
$[-\infty; 80, 5]$	$18, 0$
$[80, 5; 526, 5]$	$31, 6$
$[526, 5; 1414, 5]$	$26, 9$
$[1414, 5; \infty]$	$23, 6$

Tabelle 23: Äquifrequente Diskretisierung in 4 Intervalle

Intervall (Überlebenszeit in Tagen)	**Anteil in %**
$[-\infty; 49, 5]$	$12, 6$
$[49, 5; 313, 5]$	$23, 8$
$[313, 5; 801, 5]$	$25, 8$
$[801, 5; 1673]$	$18, 9$
$[1673 : \infty]$	$19, 0$

Tabelle 24: Äquifrequente Diskretisierung in 5 Intervalle

Intervall (Überlebenszeit in Tagen)	**Anteil in %**
$[-\infty; 22, 5]$	$5, 1$
$[22, 5; 49, 5]$	$7, 4$
$[49, 5; 137, 5]$	$11, 2$
$[137, 5; 313, 5]$	$12, 6$
$[313, 5; 526, 5]$	$13, 2$
$[526, 5; 801, 5]$	$12, 6$
$[801, 5; 1161]$	$9, 1$
$[1161; 1673]$	$9, 8$
$[1673; 2334]$	$7, 7$
$[2334; \infty]$	$11, 3$

Tabelle 25: Äquifrequente Diskretisierung in 10 Intervalle

A.5. Verteilung Äquidistante Diskretisierung

Alle Werte sind ebenfalls auf eine Nachkommastellen gerundet.

Intervall (Überlebenszeit in Tagen)	**Anteil in %**
$[-\infty; 4342, 3]$	$98, 9$
$[4342, 3; 8684, 7]$	$1, 1$
$[8684, 7; \infty]$	0

Tabelle 26: Äquidistante Diskretisierung in 3 Intervalle

Intervall (Überlebenszeit in Tagen)	**Anteil in %**
$[-\infty; 3256, 8]$	$96, 3$
$[3256, 8; 6513, 5]$	$3, 6$
$[6513, 5; 9770, 3]$	$0, 1$
$[9770, 3; \infty]$	0

Tabelle 27: Äquidistante Diskretisierung in 4 Intervalle

Intervall (Überlebenszeit in Tagen)	**Anteil in %**
$[-\infty; 2605, 4]$	$93, 0$
$[2605, 4; 5210, 8]$	$6, 7$
$[5210, 8; 7816, 2]$	$0, 3$
$[7816, 2; 10421, 6]$	0
$[10421, 6; \infty]$	0

Tabelle 28: Äquidistante Diskretisierung in 5 Intervalle

Intervall (Überlebenszeit in Tagen)	**Anteil in %**
$[-\infty; 1302, 7]$	$74, 6$
$[1302, 7; 2605, 4]$	$18, 4$
$[2605, 4; 3908, 1]$	$5, 4$
$[3908, 1; 5210, 8]$	$1, 2$
$[5210, 8; 6513, 5]$	$0, 2$
$[6513, 5; 7816, 2]$	$0, 1$
$[7816, 2; 9118, 9]$	0
$[9118, 9; 10421, 6]$	0
$[10421, 6; 11724, 3]$	0
$[11724, 3; \infty]$	0

Tabelle 29: Äquidistante Diskretisierung in 10 Intervalle